Exploratory Data Analysis (EDA)

Titanic Dataset Case Study

By Awais Manzoor

Goal

Understand survival drivers.

Data

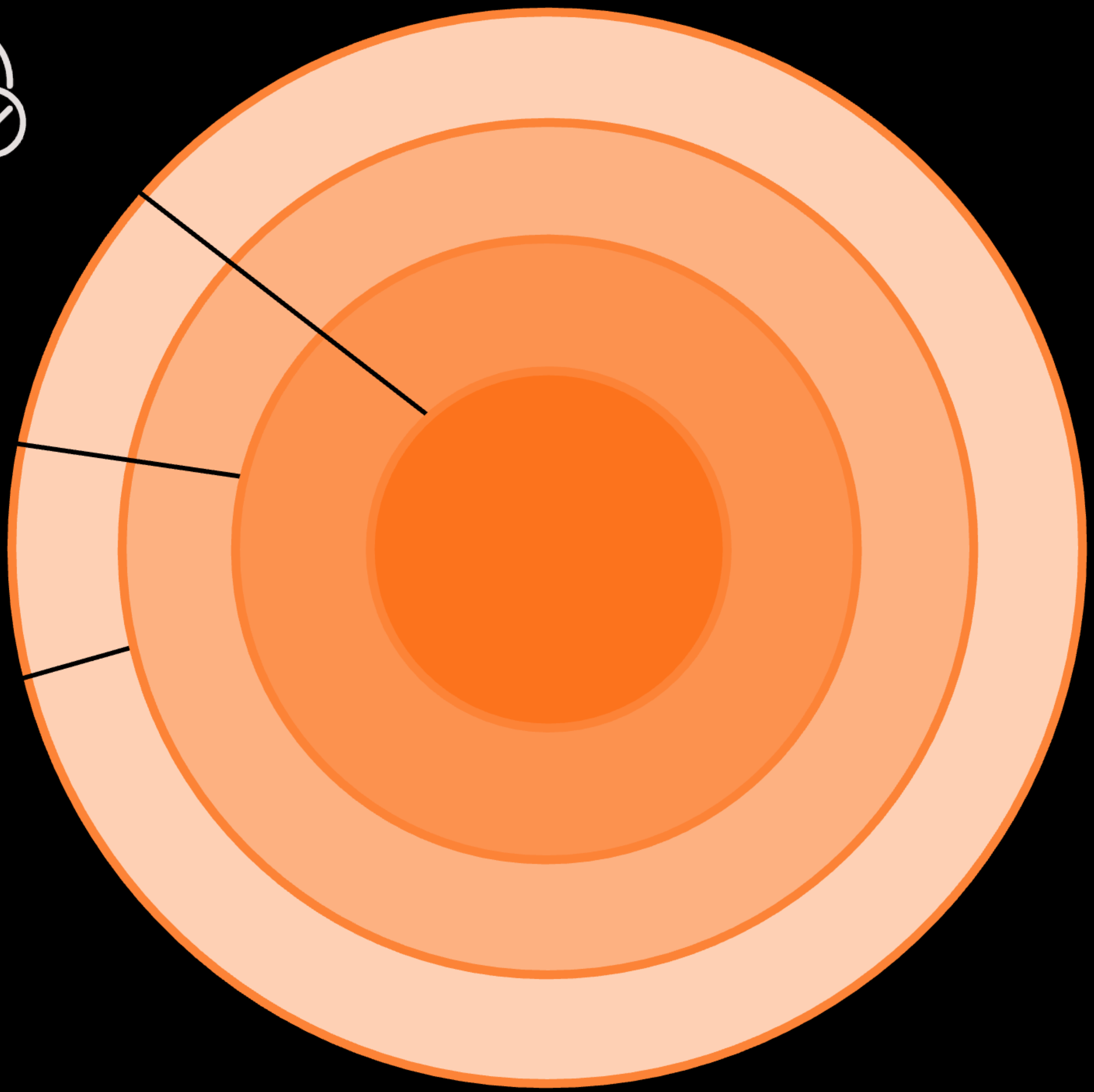
Passengers, features, and labels.

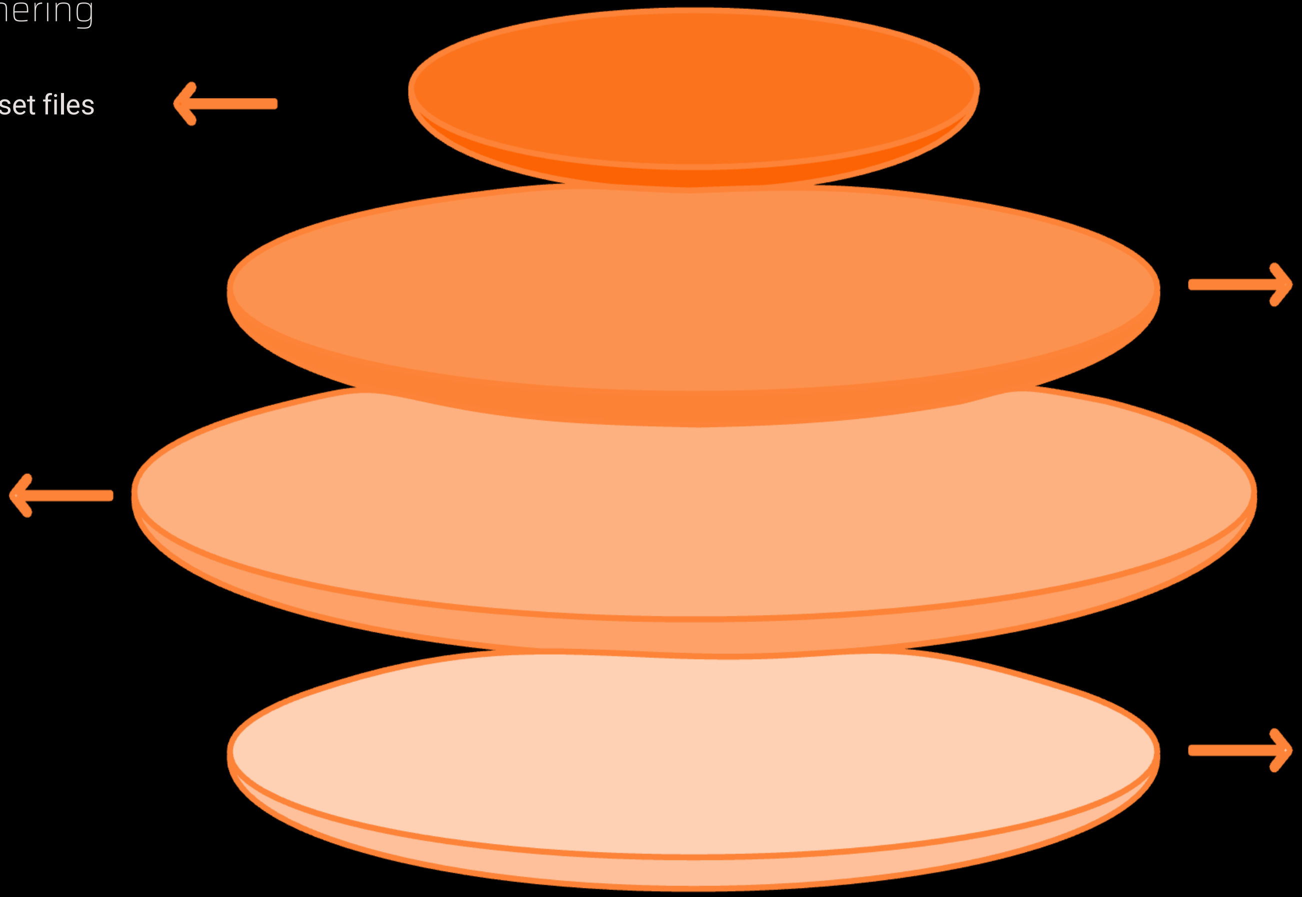
Analysis

Summary stats and visualizations.

Insights

Key predictors and patterns.





Data Gathering

Collect Titanic dataset files

Data Cleaning

Handle missing values and formats

Feature Analysis

Explore relationships and distributions

Model Insights

Summarize findings and implications

What is Exploratory Data Analysis (EDA)?

Summarize Main Characteristics

A process of analyzing datasets to summarize their main characteristics using visual and statistical methods.

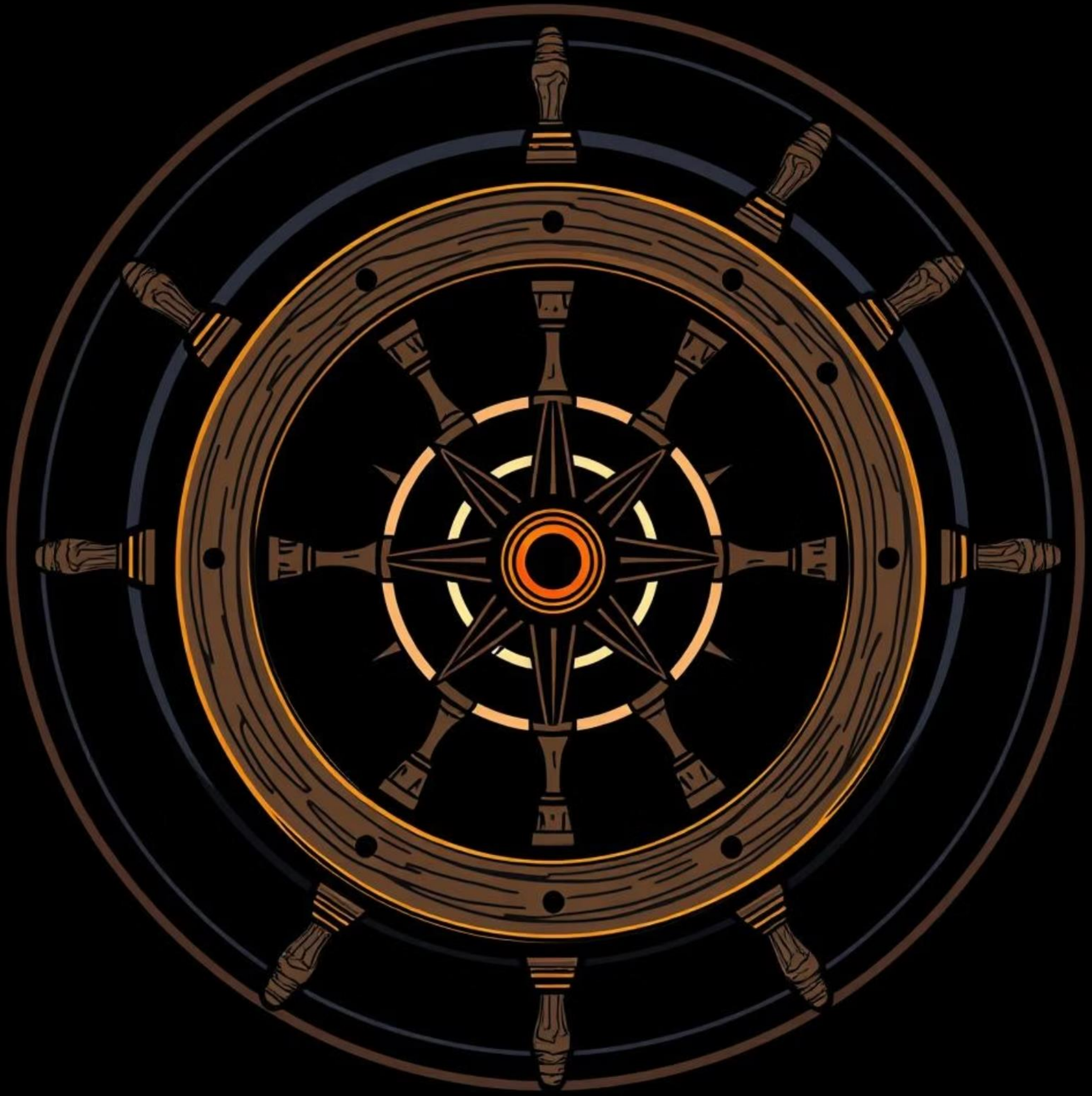
Uncover Patterns & Anomalies

Helps in identifying hidden patterns, spotting unusual data points, and validating initial assumptions.

Data Cleaning & Visualization

Involves critical steps like data cleaning, robust visualization, and preliminary statistical analysis.

Why the Titanic Dataset?



- **Classic Learning Resource:** An iconic dataset widely used for introducing data analysis and machine learning concepts.
- **Mixed Data Types:** Contains a rich blend of numerical (e.g., Age, Fare) and categorical variables (e.g., Sex, Pclass), offering diverse analysis opportunities.
- **Clear Objective:** The primary goal is to understand the factors that influenced survival, making it an intuitive problem to tackle.

Steps in Performing EDA

01

Data Collection

Gathering relevant data from various sources.

02

Data Cleaning

Addressing missing values, handling duplicates, and correcting data inconsistencies.

03

Univariate Analysis

Analyzing individual variables to understand their distributions.

04

Bivariate Analysis

Exploring relationships between pairs of variables.

05

Feature Engineering

Creating new variables from existing ones to improve model performance.

06

Key Insights & Visualization

Summarizing findings and presenting them through compelling data visualizations.

Titanic Dataset Overview

The Titanic dataset contains comprehensive information about each passenger on the ill-fated voyage.

- **Passenger Details:** Includes 'Age', 'Sex', 'Pclass' (Passenger Class), 'Fare', 'Cabin', and 'Embarked' (Port of Embarkation).
- **Target Variable:** The crucial 'Survived' column (0 = No, 1 = Yes) is what we aim to predict.
- **Use Case:** This dataset is ideal for binary classification tasks in machine learning.

Univariate Analysis: A Glimpse into Passenger Demographics

Age Distribution

Most passengers fell within the 20-40 age bracket, indicating a predominantly adult population on board.

Fare Distribution

The fare distribution was heavily right-skewed with significant outliers, showing a few passengers paid very high prices.

Gender Imbalance

The passenger list showed a clear majority of males compared to females.

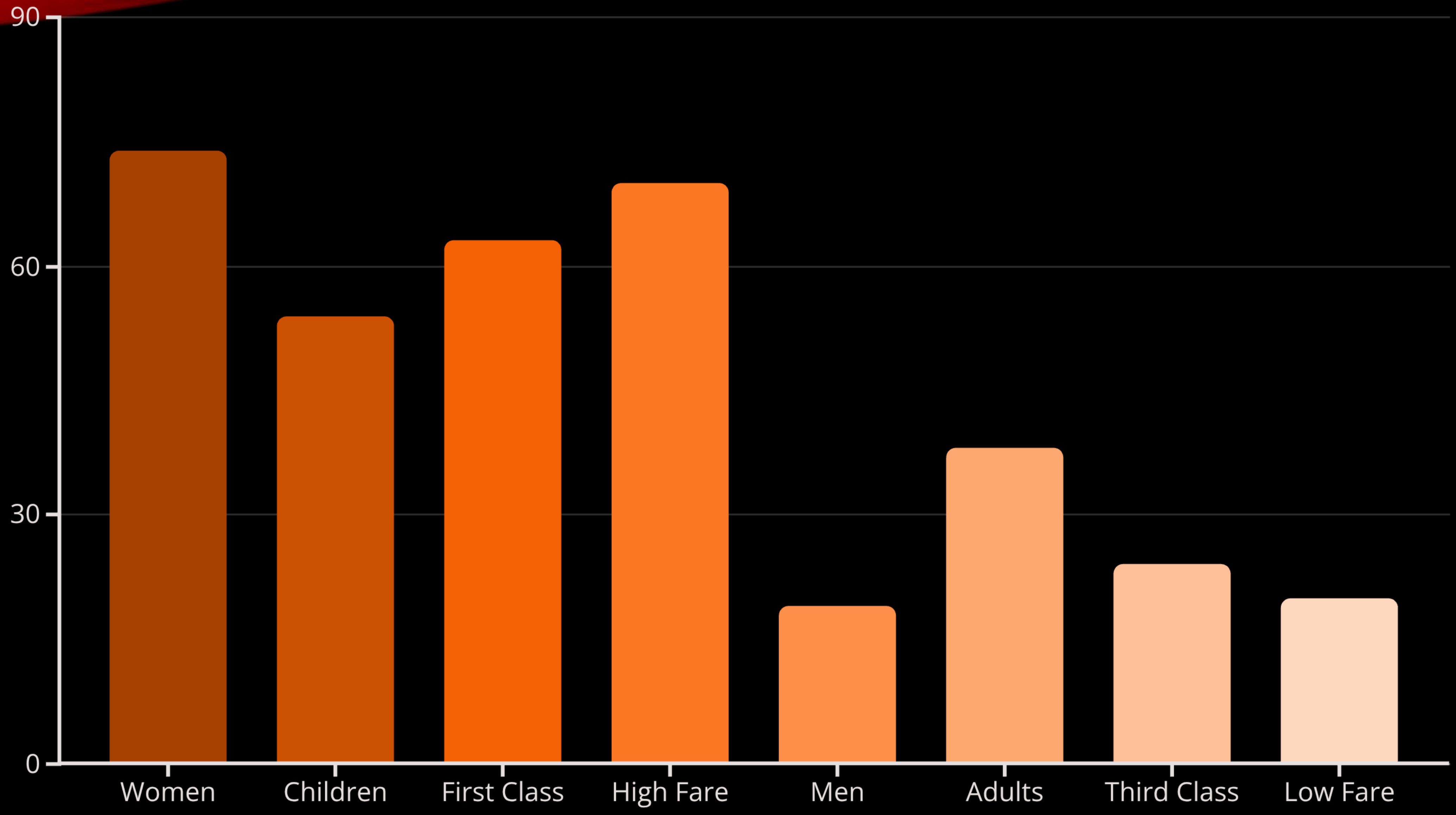
Class Distribution

A substantial portion of passengers were traveling in Class 3, reflecting the ship's diverse socio-economic cross-section.

Mixed Features

`SibSp`	– Number of siblings/spouses aboard
`Parch`	– Number of parents/children aboard
`Ticket`	– Ticket number (alphanumeric)
`Cabin`	– Cabin number (alphanumeric, missing for many)
`Name`	– Passenger’s name

Bivariate Analysis: Survival Factors



Analysis reveals significant correlations: Women and children, along with First-Class passengers, had substantially higher survival rates. A higher fare also correlated positively with survival.

Feature Engineering Ideas

1

Family Size

Combine 'SibSp' (siblings/spouse) and 'Parch' (parents/children) with 1 (for self) to create a 'FamilySize' variable. This can reveal if traveling with family impacted survival.

2

Individual Fare

Divide 'Fare' by 'FamilySize' to calculate the fare paid per person. This normalizes the fare and might better reflect individual economic status.

3

Title Extraction

Extract titles (e.g., Mr., Mrs., Miss, Master, Dr.) from the 'Name' column. Titles often convey social status, age, or marital status, which can be highly predictive.

4

Family Survival Patterns

Group passengers by 'Surname' to identify families. Then, analyze survival rates within families, as the survival of one member might influence others.

Key Insights from Titanic EDA

The exploratory data analysis of the Titanic dataset revealed several critical factors influencing survival:

1

Gender & Age Priority

Women and children had significantly higher chances of survival, likely due to the "women and children first" protocol.

2

Socio-Economic Status

First Class passengers were disproportionately more likely to survive compared to Third Class passengers, highlighting socio-economic disparity.

3

Fare & Survival

A higher fare paid for a ticket correlated positively with a better survival probability.

4

Enhanced Features

Creating engineered features such as 'FamilySize' and 'Title' can significantly improve the predictive power of future models.

Conclusion & Next Steps

ⓘ EDA: Uncovering Hidden Narratives

Exploratory Data Analysis is a powerful first step in any data science project, enabling us to uncover hidden insights and understand the data's underlying structure.

✓ Titanic: A Perfect Playground

The Titanic dataset serves as an excellent, accessible resource for practicing and mastering fundamental data analysis techniques.

Your Journey Continues...

Now that we've explored the data, the logical next step is to build predictive models using these insights to forecast survival outcomes.

💡 Share your thoughts on these findings or **try your own EDA** on the Titanic dataset!