Task 5: Exploratory Data Analysis (EDA) Report on the Titanic Dataset

Objective: Extract insights regarding survival patterns, trends, and anomalies using visual and statistical exploration.

1. Initial Data Summary and Anomalies

The initial statistical exploration provided the foundational understanding of the dataset's structure, which is crucial for data preparation.

Feature	Insight	Anomaly/Cleaning Need
Data Size	891 entries (passengers)	None
Survival Rate	Mean of Survived is 0.384 (38.4% survival rate).	The target variable is imbalanced, with more non-survivors than survivors.
Fare	Median is $14.45; Max is $512.33.	Extreme outliers in the upper range (high fares).
Missing Data	Age (177 missing), Cabin (687 missing), Embarked (2 missing).	The cabin is highly sparse. Age requires robust imputation.
Categorical Mode	Most frequent Sex is Male (577); most frequent Embarked port is 'S' (Southampton).	The 2 missing Embarked values were imputed using the mode ('S').

2. Visual Observations and Key Trends

Visual exploration identified the primary factors driving passenger survival.

A. Observations from Distribution Plots (Histograms & Boxplots)

Feature	Visual	Observation
Age	Histogram	Distribution is slightly right-skewed, with the majority of passengers being young adults (20s-30s).
Fare	Boxplot	Confirmed the presence of significant outliers in the high fare range, indicating a small number of very wealthy passengers.

B. Identification of Relationships and Trends

The most critical factor affecting survival was the **social hierarchy** on the ship.

Relationship	Visual	Trend/Insight
Survival by Sex	Bar Plot	Females had a dramatically higher survival rate (~74%) than Males (~19%).
Survival by Pclass	Bar Plot	Survival is highest in 1st Class (~63%) and lowest in 3rd Class (~24%), showing a clear trend where socioeconomic status correlates with survival probability.

C. Correlation Analysis (`sns.heatmap()` & `sns.pairplot()`)

The correlation heatmap (of numerical features) and pairplot provided statistical backing for the observed trends.

Plot	Relationship	Correlation (r) / Trend
Heatmap	Survived vs. Pclass	−0.34 (Strong Negative). Confirms 1st Class passengers (low Pclass value) were more likely to survive.
Heatmap	Survived vs. Fare	0.26 (Moderate Positive). Higher fares correlate with higher survival, serving as a proxy for Pclass.

3. Summary of Findings

The Exploratory Data Analysis on the Titanic dataset has established the following patterns, trends, and anomalies:

1. **Dominant Survival Pattern (Gender Bias):** Survival was overwhelmingly influenced by **gender**, indicating that the "women and children first" protocol was effectively enforced.
2. **Socioeconomic Trend: Passenger Class** is the second most crucial predictor. Passengers with higher socioeconomic status (1st Class) had a significantly greater chance of survival than those in 2nd or 3rd Class. This is a direct trend of **privilege and access to resources** (e.g., lifeboats).
3. **Data Anomalies:** The high volume of missing data in the `Cabin` column and the extreme **outliers** in the `Fare` distribution represent challenges that must be addressed (through feature engineering or removal) before the dataset is used for predictive modeling.