EXPLORATORY DATA ANALYSIS OF THE TITANIC DATASET: REPORT OF FINDINGS

Objective: To analyze the Titanic passenger dataset to identify key patterns, trends, and relationships that influenced survival outcomes.

1. Initial Data Inspection

The first step was to get a high-level overview of the dataset's structure and contents.

- **Structure**: The dataset contains **891 passenger records** and **12 columns**.
- **Data Types**: The data is a mix of numerical (Age, Fare, Pclass) and categorical (Sex, Embarked) information.
- **Missing Data**: A critical initial finding was the presence of significant missing values, particularly in the Age and Cabin columns, which must be accounted for in any deeper analysis.

A statistical summary of the numerical data revealed:

- **Survival Rate**: Approximately **38%** of the passengers in this dataset survived.
- **Passenger Age**: The average age was about **30 years**, ranging from infants to an 80-year-old.
- **Fare**: Ticket prices were heavily skewed, with most passengers paying a low fare, but a few paying a significantly higher amount.

---

2. Visual Analysis and Key Findings

Visualizations were crucial for uncovering the relationships between different passenger attributes and their survival.

Histogram of Passenger Age

- **Observation:** The age distribution shows that the majority of passengers were young adults between 20 and 40 years old. There was also a notable peak for young children, highlighting a key demographic group.

Boxplot of Age by Passenger Class

- **Observation:** A clear trend emerges: 1st class passengers were, on average, older than 2nd class passengers, who were in turn older than 3rd class passengers. This indicates a correlation between wealth/status and age on the vessel.

Scatterplot of Age vs. Fare

- **Observation:** This plot shows no strong linear relationship between age and the fare paid. However, when coloured by survival status, it becomes evident that survivors (orange dots) are more concentrated in the higher-fare regions, regardless of age.

Correlation Heatmap

- **Observation:** The heatmap provides a quick numerical summary of relationships. The strongest correlations related to survival were:
  - **Passenger Class (Pclass):** A negative correlation of **-0.34**, meaning a lower class number (e.g., 1st class) was associated with a higher survival chance.
  - **Fare:** A positive correlation of **0.26**, confirming that paying a higher fare was linked to a better survival outcome.

Bar Plots: Survival Rate by Social Factors

- **Observation:** These plots reveal the most critical findings:
  - **By Sex:** Gender was a decisive factor. Females had a survival rate of nearly **75%**, whereas males had a rate below **20%**.
  - **By Class:** Social status was a primary determinant. Over **60%** of 1st class passengers survived, compared to less than **25%** of 3rd class passengers.

---

3. Summary of Findings

The exploratory data analysis of the Titanic dataset leads to a clear and powerful conclusion: **a passenger's chance of survival was not determined by luck, but by their social standing and gender.**

The three main takeaways are:

1. **Social Class Was a Major Predictor of Survival:** Wealth and status, represented by passenger class, provided a significant survival advantage.
2. **"Women and Children First" Was a Real Policy:** The data overwhelmingly shows that women and young children were given priority during the evacuation.
3. **Demographics Were Not Uniform:** The passenger population was not evenly distributed; it was predominantly young and male, with most passengers traveling in 3rd class.In essence, the data tells a human story of hierarchy and social norms prevailing even in the face of disaster. A person's identity—their wealth, their gender, their age—was the most important factor in determining whether they lived or died.