Week 2 Report: Titanic Data Preprocessing & Insights

1. Project Overview

This project involves cleaning and analyzing the Titanic dataset to identify factors that influenced passenger survival. The goal was to transform raw, messy data into a format suitable for analysis and visualize key trends.

2. Data Preprocessing Steps (The "Cleaning")

To ensure the data was accurate and ready for visualization, the following steps were taken:

- **Handling Missing Values:** The "Age" column had several missing entries. I used **Mean Imputation** (filling gaps with the average age of ~29) to avoid losing valuable passenger records.
- **Categorical Encoding:** Since machines only understand numbers, the "Sex" column was encoded: **Female = 0** and **Male = 1**.
- **Normalization:** The "Fare" column had a wide range (from $0 to $512). I applied **Min-Max Scaling** to normalize these values between 0 and 1, making it easier to compare distributions.

3. Key Visualizations & Findings

Based on the 4-5 charts created in the Jupyter Notebook, here are the primary findings:

- **Gender and Survival:** The most significant insight was the survival gap. Females had a much higher survival rate (over 70%) compared to males (under 20%).
- **Class Priority:** Passengers in **1st Class (Pclass 1)** had the highest survival percentage. This suggests that socio-economic status played a major role in lifeboat access.
- **Age Distribution:** The visualization showed a "spike" in survival for children (ages 0-10). Conversely, the highest mortality was seen in men aged 20-35.
- **Fare Correlation:** There was a positive correlation between higher fares and survival rates, confirming that those who paid more were more likely to survive.

4. Actionable Insights for "Educators" (Stakeholders)

If we were to use this data for safety training or historical analysis, the insights would be:

1. **Safety Protocol Impact:** The "Women and Children First" protocol was clearly followed and effective in saving those specific groups.

2. **Structural Inequality:** The data highlights a need for equalized emergency access, as lower-class cabins (3rd Class) faced much higher risks.
3. **Resource Allocation:** In emergency scenarios, the data shows that proximity to lifeboats (linked to cabin class) is the most critical factor for survival.