Dataset Analysis Report: Titanic Dataset

1. Dataset Overview

- **Dataset Name:** Titanic - Machine Learning from Disaster.
- **Source:** Loaded via Pandas from remote repository (DataScienceDojo).
- **Size:** The dataset consists of **891 rows** (passengers) and **12 columns** (features).
- **Suitability for ML:** Yes, the dataset is suitable for Supervised Learning (Classification) as it is labelled data with sufficient rows to train a basic model.

2. Data Types & Structure

- **Numerical Features:** Age (Continuous), Fare (Continuous), SibSp (Discrete), Parch (Discrete).
- **Categorical Features:** Sex (Nominal), Embarked (Nominal), Name (Nominal), Ticket (Nominal).
- **Ordinal Features:** Pclass (1st, 2nd, 3rd class implies an order/hierarchy).
- **Binary/Target Variable:** Survived (0 = No, 1 = Yes).

3. Statistical Observations

- **Age:** The average passenger age is approx. 29.7 years. The youngest is an infant (0.42) and the oldest is 80.
- **Fare:** There is a high variance in ticket price, ranging from 0 to 512, indicating a disparity in passenger wealth.

4. Data Quality & Observations

- **Missing Values:**
    - Cabin: High volume of missing data (687 missing). This column might need to be dropped.
    - Age: 177 missing values. These will need imputation (filling with mean/median) before modelling.
    - Embarked: Only 2 missing values.
- **Data Imbalance:**
    - The dataset is slightly imbalanced. 61.6% of passengers did not survive, while 38.3% survived. Accuracy alone may not be the best metric; F1-score might be needed.