Dataset Card — Titanic

Name: Titanic - Machine Learning from Disaster
Source: Kaggle Titanic Dataset
Description: Passenger data from the RMS Titanic's ill-fated 1912 voyage, used to predict survival outcomes based on demographics, ticket class, and other features.
Size: ~891 rows × 12 columns (train set)
Columns:

Column	Type	Description
PassengerId	int	Passenger identifier
Survived	int (0/1)	Survival status (1 = survived)
Pclass	int	Ticket class (1 = Upper, 2 = Middle, 3 = Lower)
Name	string	Passenger name
Sex	string	Gender
Age	float	Passenger age
SibSp	int	Number of siblings/spouses aboard
Parch	int	Number of parents/children aboard
Ticket	string	Ticket number
Fare	float	Passenger fare
Cabin	string	Cabin number
Embarked	string	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Titanic Dataset — EDA & Preprocessing Roadmap

---

1 Understand the Dataset

- Each row = **one passenger** on the Titanic.
- **Columns:**
    - PassengerId (identifier)
    - Survived (target: 0 = no, 1 = yes)
    - Pclass (ticket class: 1st, 2nd, 3rd)
    - Name (text)

    - Sex (categorical)
    - Age (numeric)
    - SibSp (number of siblings/spouses aboard)
    - Parch (number of parents/children aboard)
    - Ticket (text)
    - Fare (numeric)
    - Cabin (text)
    - Embarked (port of embarkation: C, Q, S)
- **Goal:** Explore factors affecting survival.

---

2 Initial Data Inspection

- Shape of dataset (rows × columns).
- First 5 rows (head).
- Column data types.
- Identify numeric vs categorical columns.
- Count missing values per column.
- Summary statistics for numeric columns (mean, median, std, min, max).

---

3 Preprocessing Steps

A. Handle Missing Values

- **Age** → fill with median or group median (by Pclass & Sex).
- **Cabin** → many missing; consider new column Has_Cabin or drop.
- **Embarked** → fill with most common port (mode).
- **Fare** → fill with median if missing.

B. Handle Duplicates

- Rare in Titanic dataset, but check PassengerId or Name duplicates.

C. Feature Engineering

- **Title** from Name (Mr, Mrs, Miss, Master, etc.).
- **FamilySize** = SibSp + Parch + 1.
- **IsAlone** = 1 if FamilySize = 1 else 0.
- **AgeGroup**: Child, Teen, Adult, Senior.

- **FareBin**: Low, Medium, High fare categories.
- **Deck** from first letter of Cabin.