

Observations:

- The dataset has **891 rows** and **12 columns**.
- Columns **Age**, **Cabin**, and **Embarked** have missing values.
- Age has some missing data (~177 missing values).
- Fare values are highly spread out (min = 0, max = 512).
- Around **550 passengers did not survive**, and about **340 survived**.
- Most passengers belonged to **3rd class** (Pclass = 3).
- **Age distribution** is **slightly right-skewed** — most passengers are between **20-40 years** old.
- There are **outliers in Fare** — few passengers paid very high fares.
- **Females** had a **much higher survival rate** than males.
- **1st Class passengers** had **highest survival rates**, followed by 2nd class.
- Younger passengers and passengers paying **higher fares** had better survival chances.
- **Fare** and **Pclass** are **strongly negatively correlated** (Higher class = Higher fare).
- **Survived** is **positively correlated with Fare** and **negatively correlated with Pclass**.
- No strong multicollinearity among features.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton