TITANIC DATASET
Exploratory Data Analysis (EDA) Report

Data Analyst Internship — Task 5

This report presents a comprehensive exploratory data analysis of the **Titanic passenger dataset (train.csv — 891 records)**. Using statistical summaries, univariate, bivariate, and multivariate visualizations, we uncover the key patterns that determined passenger survival.

Tools used: Python · Pandas · Matplotlib · Seaborn

1. Dataset Overview

The Titanic training dataset contains **891 rows and 12 columns**. The key features are listed below along with data types and missing value counts.

Column	Type	Non-Null	Missing	Description
PassengerId	int	891	0	Unique passenger ID
Survived	int	891	0	Target: 0 = No, 1 = Yes
Pclass	int	891	0	Ticket class (1st / 2nd / 3rd)
Name	str	891	0	Full name of passenger
Sex	str	891	0	Gender (male / female)
Age	float	714	177 (19.9%)	Age in years
SibSp	int	891	0	Siblings / Spouses aboard
Parch	int	891	0	Parents / Children aboard
Ticket	str	891	0	Ticket number
Fare	float	891	0	Passenger fare (£)
Cabin	str	204	687 (77.1%)	Cabin number
Embarked	str	889	2	Port: S / C / Q

Descriptive Statistics (Numerical Columns):

Statistic	Survived	Pclass	Age	SibSp	Parch	Fare
Count	891	891	714	891	891	891
Mean	0.38	2.31	29.70	0.52	0.38	32.20
Std Dev	0.49	0.84	14.53	1.10	0.81	49.69
Min	0.00	1.00	0.42	0.00	0.00	0.00
25th %ile	0.00	2.00	20.12	0.00	0.00	7.91
Median	0.00	3.00	28.00	0.00	0.00	14.45
75th %ile	1.00	3.00	38.00	1.00	0.00	31.00
Max	1.00	3.00	80.00	8.00	6.00	512.33

2. Missing Value Analysis

Three columns have missing data. **Cabin (77.1%)** is largely unusable without imputation. **Age (19.9%)** requires median/model-based imputation before modelling. **Embarked** has only 2 missing values and can be filled with the mode ('S').

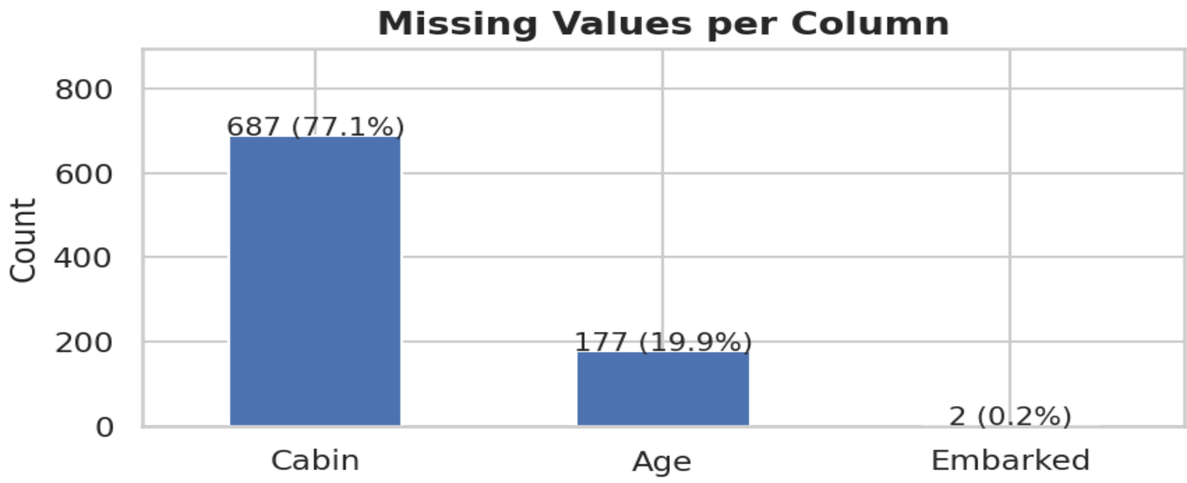


Figure 1 — Missing values per column

Observation: *Cabin has the most missingness (687/891 = 77.1%). Age is missing for 177 passengers (~20%). Both Embarked and Fare are nearly complete. Cabin should be dropped or engineered (has_cabin flag) before building a model.*

3. Univariate Analysis

Univariate analysis examines each variable individually to understand its distribution, central tendency, spread, and shape.

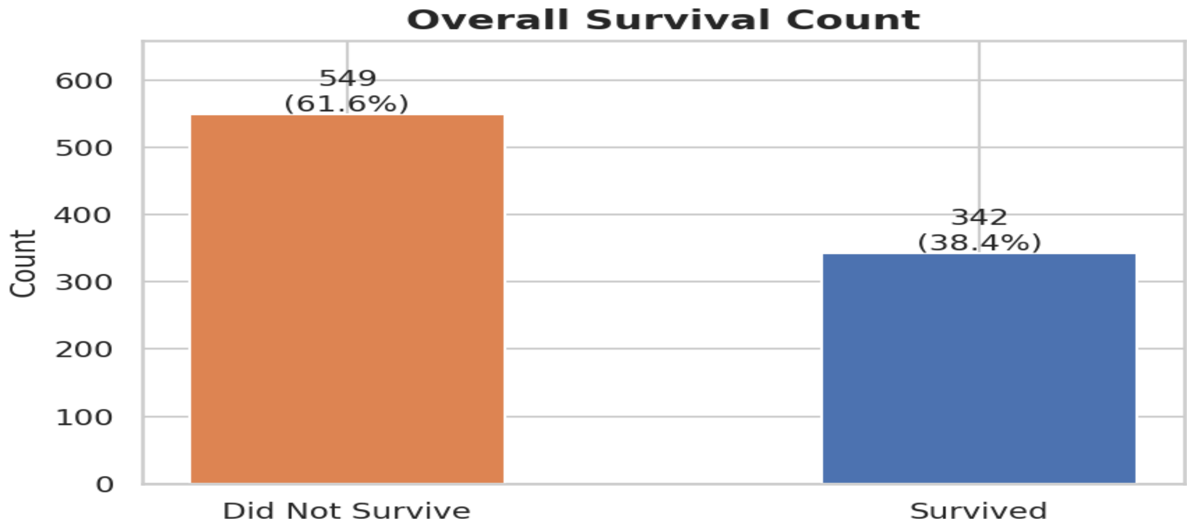


Figure 2 — Overall Survival Count

Observation: *549 passengers (61.6%) did not survive; 342 (38.4%) survived. The dataset is moderately imbalanced — stratified sampling or class weights should be used when building classification models.*

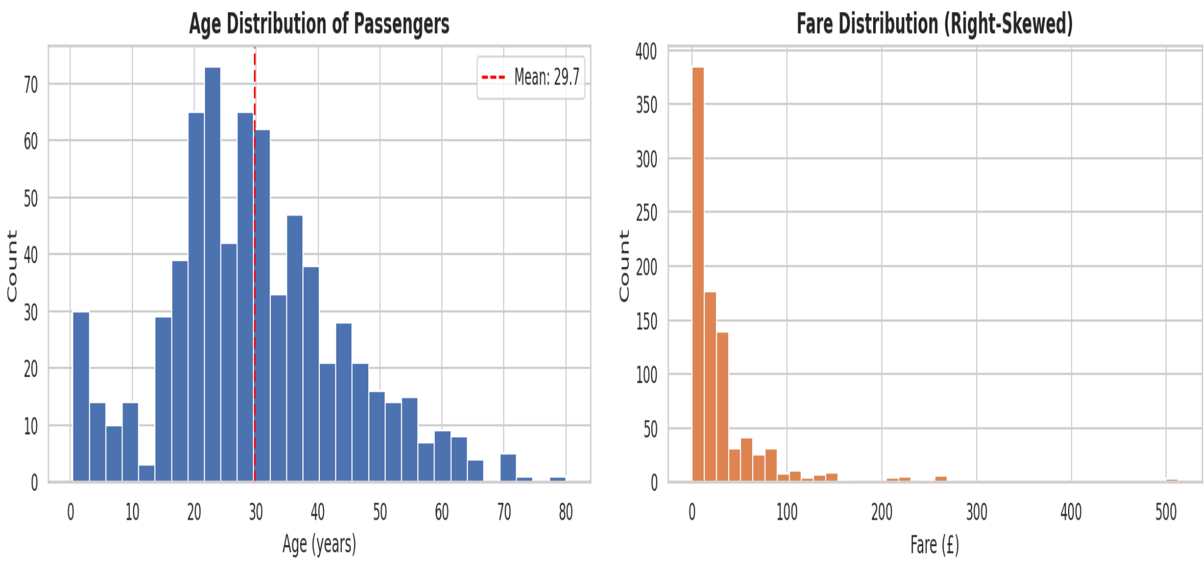
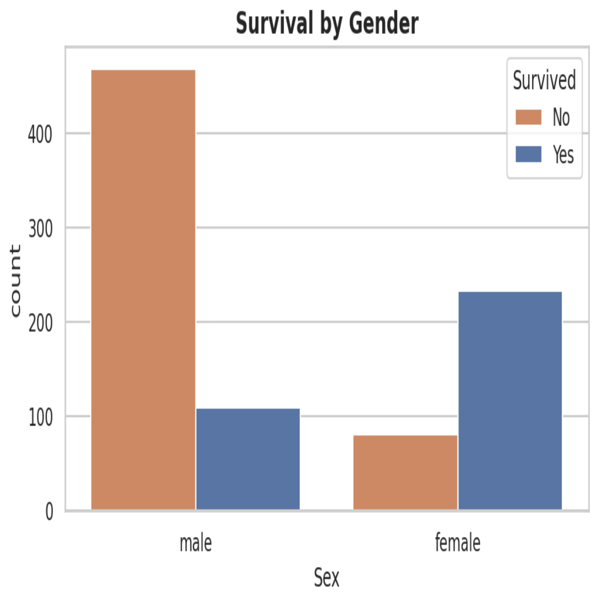


Figure 3 (left) — Age Distribution | Figure 4 (right) — Fare Distribution

Observation: *Age (left): Approximately bell-shaped with mean ~29.7 years. A small peak near 0–5 suggests children were present. 177 values are missing. Fare (right): Severely right-skewed — most passengers paid under £50 while a few paid over £500. Log transformation is strongly recommended before modelling.*

4. Bivariate Analysis

Bivariate analysis explores relationships between pairs of variables, particularly how each feature relates to the survival outcome.



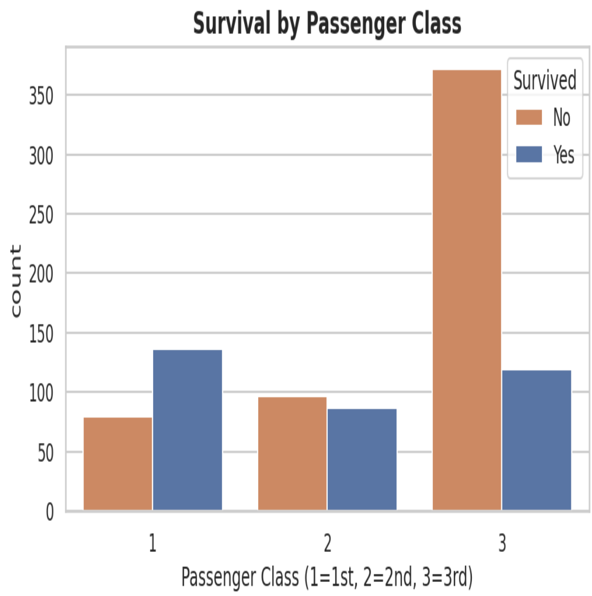
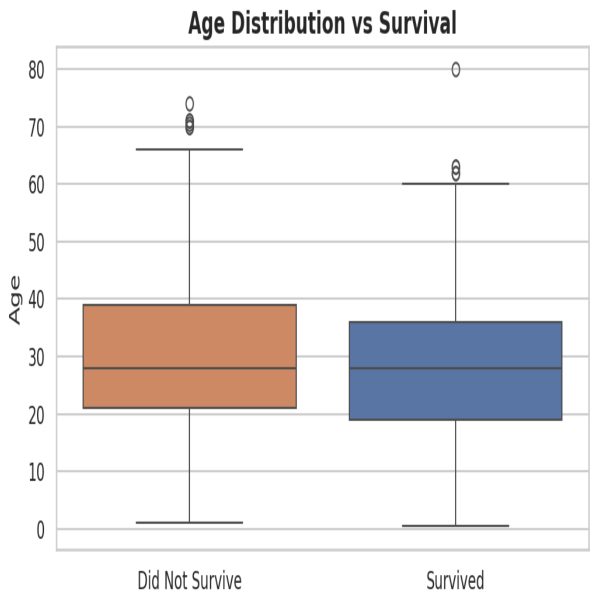


Figure 5 (left) — Survival by Gender | Figure 6 (right) — Survival by Class

Observation: *Gender (left): 233 of 314 females survived (74.2%) vs only 109 of 577 males (18.9%). Sex is the single strongest predictor of survival. Passenger Class (right): 1st class survival = 63.0%, 2nd = 47.3%, 3rd = 24.2%. Higher class correlates strongly with survival, likely due to better lifeboat access.*



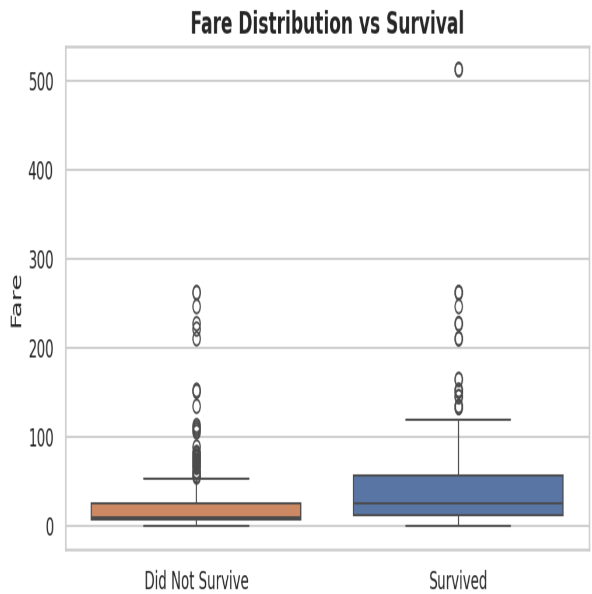


Figure 7 (left) — Age vs Survival | Figure 8 (right) — Fare vs Survival

Observation: *Age (left): Survivors had a slightly lower median age (~28) than non-survivors (~29.7). The difference is modest, but children had notably higher survival rates. Fare (right): Survivors paid a significantly higher median fare (£52 vs £22), confirming that wealthier passengers in higher classes had better survival odds.*

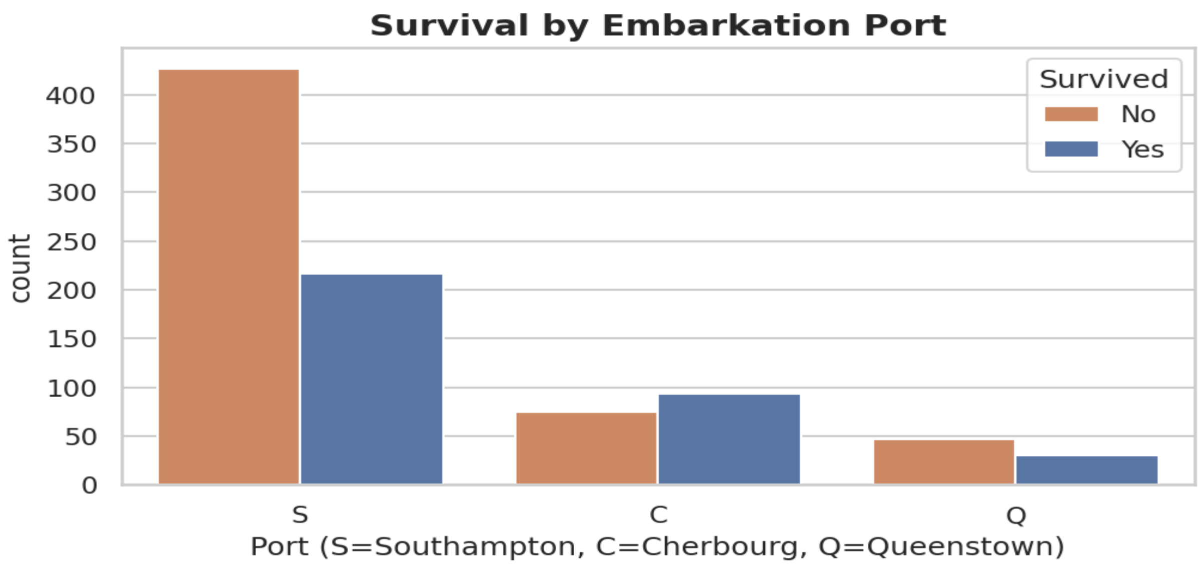


Figure 9 — Survival by Embarkation Port

Observation: *Southampton (S) was the most common boarding point (644 passengers, 72.3%) but had the lowest survival rate. Cherbourg (C) passengers had notably higher survival — likely because a greater proportion of C passengers were in 1st class. Queenstown (Q) was mostly 3rd class and had a low survival rate.*

5. Multivariate Analysis

Multivariate analysis examines interactions among three or more variables simultaneously.

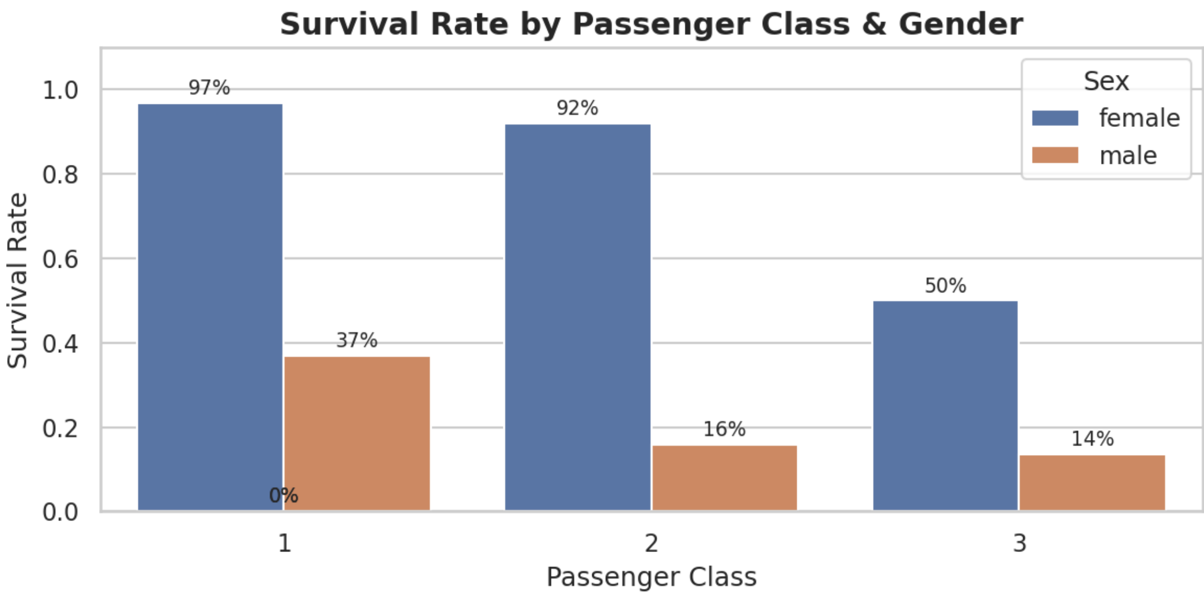


Figure 10 — Survival Rate by Passenger Class & Gender

Observation: *The interaction of Sex and Pclass reveals the full picture: female 1st-class passengers had ~97% survival; female 3rd-class ~50%. Male 1st-class: ~37%; male 3rd-class: ~14%. Gender dominates within every class. These two features together explain most of the variation in survival.*

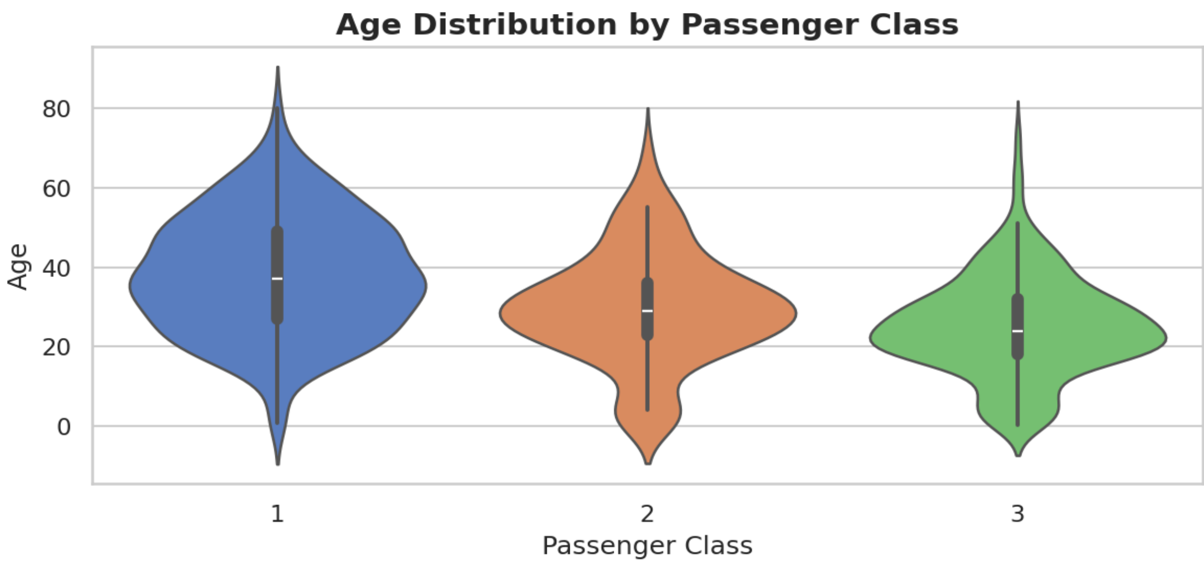


Figure 11 — Age Distribution by Passenger Class (Violin Plot)

Observation: *1st-class passengers were generally older (median ~37 years) — wealthier, established adults. 2nd-class shows a broader spread (~29 years median). 3rd-class had the youngest passengers and most variance, including many children and young migrants seeking a new life.*

6. Correlation Heatmap

The heatmap shows Pearson correlation coefficients between all numerical features. Values range from -1 (perfect negative) to +1 (perfect positive).

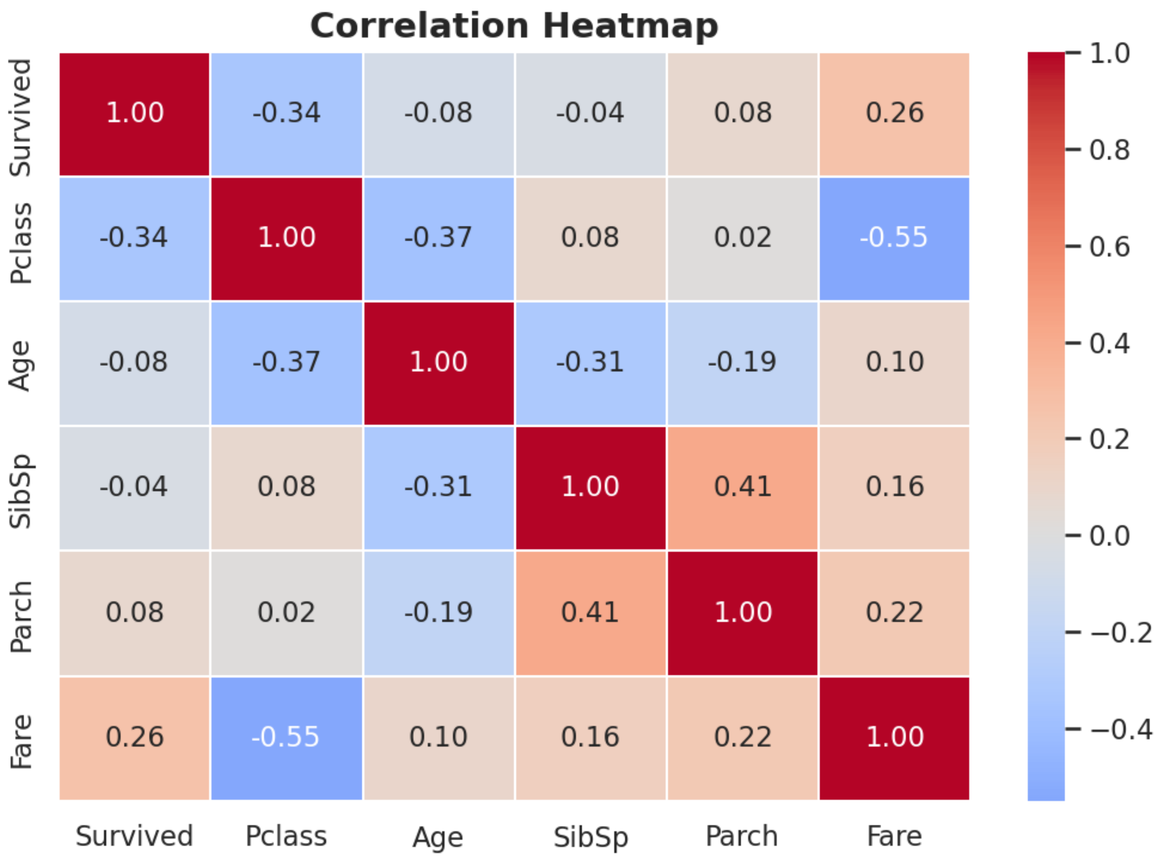


Figure 12 — Correlation Heatmap

Observation: *Key correlations with Survived: Fare (+0.26), Pclass (–0.34). Pclass and Fare are strongly negatively correlated (–0.55) — higher class means cheaper fare is inverted because class 1 = highest, class 3 = lowest, yet class 1 pays more. SibSp and Parch are slightly positively correlated (+0.41) — travelling with family. No extreme multicollinearity (|r| > 0.8) exists in the numeric features.*

7. Pairplot

The pairplot provides a visual overview of all pairwise relationships among key numerical features, with survival highlighted by colour.

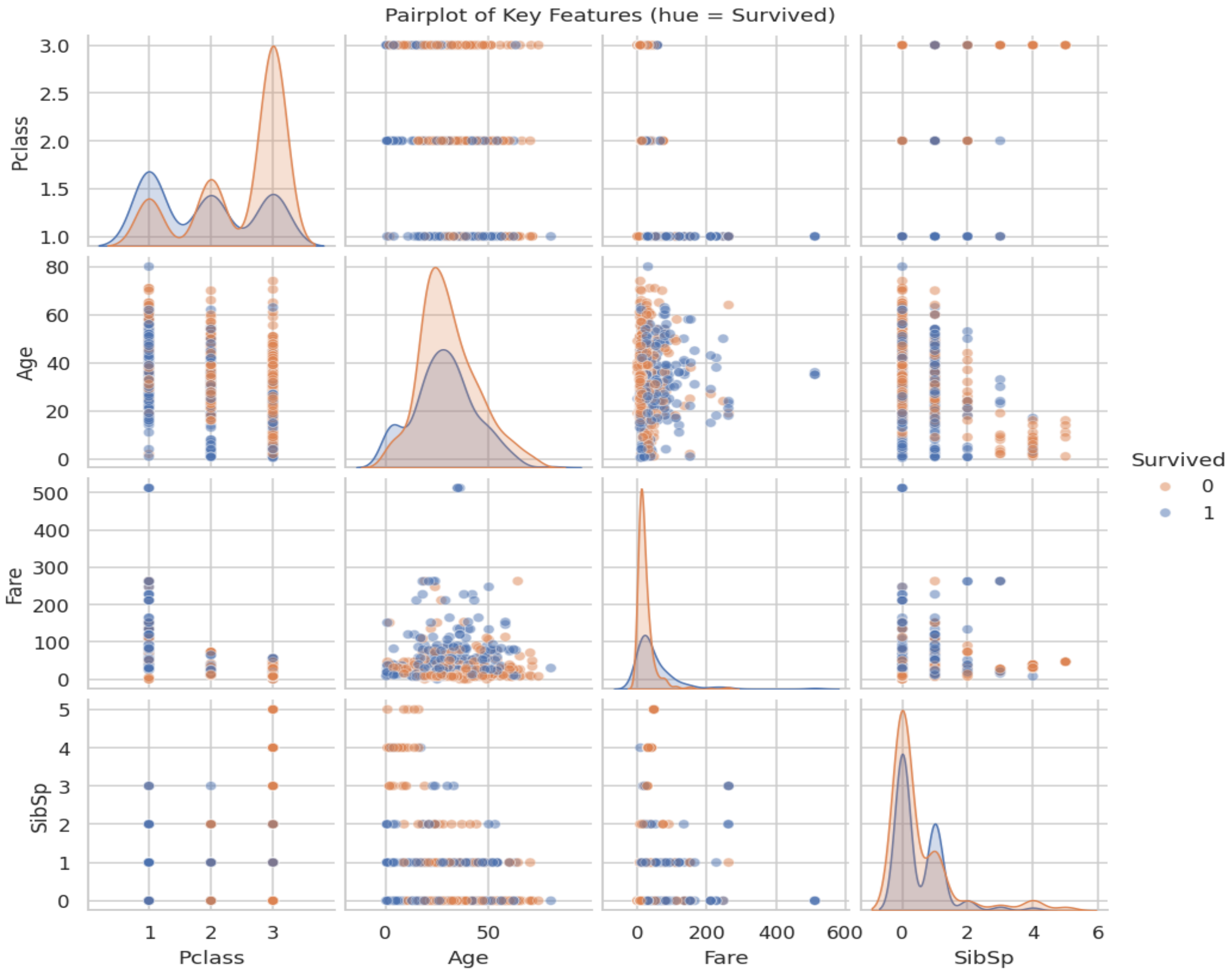


Figure 13 — Pairplot (blue = Survived, orange = Did Not Survive)

Observation: *Survivors (blue) cluster at higher Fare values and lower Pclass values across nearly all scatter plots, confirming these as key predictors. Age shows less separation — both groups span a wide age range. SibSp and Parch have limited individual predictive power but may be useful when combined with other features.*

8. Summary of Key Findings

1. Gender is the strongest survival predictor

74.2% of females survived vs 18.9% of males. The 'women and children first' protocol is clearly reflected in the data.

2. Passenger Class strongly determines survival

1st class: 63.0% | 2nd class: 47.3% | 3rd class: 24.2%. Higher-class passengers had better lifeboat access.

3. Fare and survival are positively correlated

Survivors paid a median fare of ~£52 vs ~£22 for non-survivors. Fare is a proxy for class and wealth.

4. Combined effect of Sex × Pclass is most powerful

Female 1st-class passengers had ~97% survival. Male 3rd-class: ~14%. Gender dominates within every class.

5. Age is mildly predictive

Mean age of survivors is slightly lower (~28 vs ~30). Children had higher survival rates, but the overall age effect is weak.

6. Fare is heavily right-skewed

Log transformation (np.log1p) is recommended before using Fare in ML models.

7. Age (19.9%) and Cabin (77.1%) have significant missing data

Age should be imputed (median per Pclass/Sex). Cabin should be converted to a binary 'has_cabin' feature.

8. No severe multicollinearity detected

The strongest numeric correlation is Pclass vs Fare (–0.55), which is expected and manageable.

9. Class imbalance is moderate

61.6% non-survivors vs 38.4% survivors. Use stratified train/test splits and consider class weights in models.