Titanic — Exploratory Data Analysis (Task 5)

Report of Findings
Prepared by: Kartikey Tiwari
Date: August 11, 2025

---

1. Executive summary

This EDA examines patterns in the Titanic `train.csv` dataset to understand factors associated with passenger survival. Major findings:

- **Sex** is the strongest single predictor: **females** had notably higher survival rates than males.
- **Passenger class (Pclass)** correlates with survival — **1st class** passengers survived at a much higher rate than 2nd and 3rd class.
- **Age** and **Fare** show meaningful relationships with survival: younger passengers and those who paid higher fares were more likely to survive.
- The dataset contains **missing values** (notably `Age` and `Cabin`) that require careful handling for modeling.
 Recommendations include simple imputations for `Age`, engineering features (family size, deck), and prioritizing `Sex` and `Pclass` in baseline models.

---

2. Data sources & files used

- `train.csv` — main dataset used for EDA and analysis.
- `test.csv` — reserved for later model evaluation.
- `gender_submission.csv` — sample submission file (not used for EDA).

Key columns analyzed: `PassengerId`, `Survived`, `Pclass`, `Name`, `Sex`, `Age`, `SibSp`, `Parch`, `Ticket`, `Fare`, `Cabin`, `Embarked`.

3. Objective

- Perform exploratory analysis to surface patterns, trends, and anomalies that explain survival on the Titanic.
- Generate visual and statistical insights to guide preprocessing and modeling decisions.

4. Methodology & tools

- Tools: Python, Pandas, Matplotlib, Seaborn (Jupyter Notebook deliverable).
- Steps:
  1. Data loading and initial inspection (`.info()`, `.describe()`, `.head()`).
  2. Missing value analysis.
  3. Univariate analysis (distributions and counts).
  4. Bivariate analysis (survival vs categorical/numerical features).
  5. Correlation analysis and pairwise visualizations.
  6. Summarize findings and suggest next steps.

5. Data quality & missing values

- `Age`: **~20%** missing (needs imputation — median or model-based).
- `Cabin`: heavily missing (often >70%) — not directly usable unless engineered (extract deck from cabin letter).
- `Embarked`: a small number of missing values (2–3 rows) — can impute with mode.
- `Fare`: complete in `train.csv` (or nearly complete).
  Implication: Impute `Age`, drop or engineer from `Cabin`, and fill `Embarked` with

mode.

---

6. Univariate analysis (major variables)

6.1 Survived

- Distribution: Two classes (0 = died, 1 = survived).
- Overall survival rate: **~38%** (exact percent depends on dataset run).

6.2 Sex

- Counts: More males than females on board.
- Survival: Females have a substantially higher survival proportion than males.

6.3 Pclass

- Counts: Majority in 3rd class, fewer in 1st.
- Survival: 1st class survival rate >> 2nd class > 3rd class.

6.4 Age

- Distribution: Right-skewed with many children and adults; some elderly passengers.
- Observations: Children showed relatively higher survival when compared to certain adult age groups (requires binned analysis).

6.5 Fare

- Distribution: Right-skewed; a few passengers paid very high fares.
- Higher fares generally correlate with higher survival.

6.6 Family (SibSp, Parch)

- Many passengers travel alone (low SibSp/Parch).

- Moderate relationship: having family sometimes improved survival, but larger families could be at risk.

---

7. Bivariate & multivariate findings

7.1 Sex vs. Survived

- Visualization: Countplot of `Sex` split by `Survived`.
- Finding: Females ~70% survival; males ~20% survival (approximate).

7.2 Pclass vs. Survived

- Visualization: Countplot of `Pclass` with hue `Survived`.
- Finding: 1st class survival rate substantially higher; 3rd class had the highest fatalities.

7.3 Age vs. Survived

- Visualization: Boxplots and KDEs for Age by survival.
- Finding: Survivors slightly younger on average; children had a favorable survival rate.

7.4 Fare vs. Survived

- Visualization: Violin/boxplots or scatter of Fare colored by Survived.
- Finding: Survivors tend to have higher median fares; extreme fares correspond to survivors in many cases.

7.5 Embarked vs. Survived

- Visualization: Countplot of `Embarked` with hue `Survived`.
- Finding: Slight variations by port (C/Q/S) — need statistical test to confirm significance.

7.6 Correlation matrix

- Variables: numeric-only (`Survived`, `Age`, `SibSp`, `Parch`, `Fare`, `Pclass`).
- Observations:
  - `Pclass` negatively correlates with `Fare` (higher class → higher fare).
  - `Survived` positively correlates with `Fare` and negatively with `Pclass` (because lower number = higher class).
  - Correlations are moderate — good to combine categorical and numerical features in models.