Exploratory Data Analysis (EDA) Report

Dataset: Titanic (via seaborn library)
Objective: Identify key data patterns, visualize relationships, and summarize findings to inform modeling.

1. Dataset Overview

- **Source:** Seaborn's built-in Titanic dataset (no manual download required).
- **Rows:** 891 passengers.
- **Columns:** 15 features including demographics, ticket details, and survival outcome.
- **Target variable:** `survived` (0 = died, 1 = survived).

Key columns:

- **Demographics:** `sex`, `age`, `who` (man/woman/child).
- **Travel details:** `pclass` (ticket class), `fare`, `embarked`, `embark_town`, `deck`.
- **Family context:** `sibsp` (siblings/spouses aboard), `parch` (parents/children aboard), `alone`.
- **Outcome:** `survived`, `alive`.

2. Missing Values Analysis

- **Age:** ~20% missing.
- **Deck:** ~77% missing.
- **Embark_town:** 2 missing values.
- **Other columns:** Complete.

Treatment:

- Age imputed by group median (based on `who`).
- Embark_town filled with mode (most common port).
- Deck treated as categorical with "Unknown".

3. Univariate Analysis

- **Survival rate:** ~38% survived, 62% died.
- **Age distribution:** Right-skewed; majority between 20–40 years.
- **Fare distribution:** Highly skewed; most fares < 50, few extreme outliers > 200.

4. Bivariate Analysis

- **Survival by sex:**
    - Female survival rate ~74%.

    - Male survival rate ~19%.
      → Strong gender effect.
- **Survival by class:**
    - 1st class: ~63% survived.
    - 2nd class: ~47% survived.
    - 3rd class: ~24% survived.
      → Higher class strongly correlated with survival.
- **Age vs survival:**
    - Children (<12) had higher survival.
    - Seniors (>60) had lower survival.
- **Fare vs survival:**
- Higher fares associated with higher survival (wealthier passengers more likely to survive).

5. Feature Engineering

- **Family size:** `family_size = sibsp + parch + 1`.
    - Small families (2–4) had better survival than solo travelers or very large families.
- **Is alone:** Binary indicator.
    - Alone passengers survival rate ~30%.
    - Non-alone passengers survival rate ~50%.
- **Age groups:**
- Children: highest survival (~60%).
- Seniors: lowest survival (~20%).

6. Correlation Analysis

Correlation heatmap (numeric features):

- **Survived vs pclass:** −0.31 (negative correlation; higher class number → lower survival).
- **Survived vs fare:** +0.26 (positive correlation).
- **Survived vs is_alone:** −0.20 (being alone reduces survival).
- **Survived vs age:** −0.07 (weak negative correlation).

7. Multivariate Analysis

- **Sex + Class interaction:**
    - Female 1st class passengers had the highest survival (~95%).
    - Male 3rd class passengers had the lowest survival (~15%).
- **Age vs fare scatterplot:**
- Survivors cluster among younger passengers with higher fares.
- Non-survivors cluster among older, lower-fare passengers.

8. Statistical Tests

- **Chi-square (sex vs survival):** $p < 0.001$ → Survival strongly depends on sex.

- **Chi-square (class vs survival):** $p < 0.001$ → Survival strongly depends on class.
- **T-test (age vs survival):** $p < 0.05$ → Age distributions differ significantly between survivors and non-survivors.

9. Key Insights

1. **Gender effect:** Females had much higher survival rates than males.
2. **Class effect:** 1st class passengers were far more likely to survive than 3rd class.
3. **Fare effect:** Higher fares correlated with survival, reflecting socioeconomic advantage.
4. **Family context:** Passengers with small families had better survival odds than those alone.
5. **Age effect:** Children had higher survival; seniors had lower survival.
6. **Deck missingness:** Most deck data missing; treat as categorical “Unknown” or drop.
7. **Strong predictors for modeling:** Sex, class, fare, family_size, is_alone, age_group.

10. Conclusion

The Titanic dataset reveals clear survival patterns driven by **socioeconomic status (class, fare), gender, age, and family context**. These engineered features provide strong signals for predictive modeling. Future steps include:

- Encoding categorical variables (sex, class, embark_town, deck).
- Scaling numeric features (age, fare, family_size).
- Training classification models (Logistic Regression, Random Forest).
- Evaluating performance with accuracy, precision, recall, F1-score, and ROC-AUC.