Exploratory Data Analysis Report: Titanic Dataset

1. Dataset Overview

- **Size**: 891 passengers, 15 columns (e.g., survived, pclass, sex, age, fare, embarked).
- **Missing Values**: age (177), cabin (687), embarked (2).
- **Summary Statistics** (numerical columns):

Statistic	survived	pclass	age	sibsp	parch	fare
count	891	891	714	891	891	891
mean	0.3838	2.3086	29.699	0.5230	0.3816	32.204
std	0.4866	0.8361	14.526	1.1027	0.8061	49.693
min	0.0	1.0	0.42	0.0	0.0	0.0
25%	0.0	2.0	20.125	0.0	0.0	7.910
50%	0.0	3.0	28.0	0.0	0.0	14.454
75%	1.0	3.0	38.0	1.0	0.0	31.0
max	1.0	3.0	80.0	8.0	6.0	512.329

- **Categorical Counts**:
    - survived: 61.6% died (549), 38.4% survived (342).
    - sex: 64.8% male (577), 35.2% female (314).
    - pclass: 55.1% 3rd (491), 24.2% 1st (216), 20.7% 2nd (184).
    - embarked: 72.4% S (644), 18.9% C (168), 8.7% Q (77).

2. Visualizations and Observations

- **Pairplot**: Higher fares and lower pclass (1st class) correlate with survival. Children (age < 10) show higher survival rates.
- **Correlation Heatmap**: Notable correlations:
    - pclass vs. fare: -0.55 (higher class, higher fare).
    - survived vs. fare: 0.26 (higher fares linked to survival).
    - sibsp vs. parch: 0.41 (family sizes related).
- **Age Histogram**: Right-skewed, most passengers aged 20-40 (mean ~29.7). Few infants and elderly.
- **Fare by Survival (Boxplot)**: Survivors paid higher fares (median ~26 vs ~10). High-fare outliers among survivors.
- **Age vs. Fare (Scatterplot)**: No strong linear trend, but survivors are more common at higher fares across ages.
- **Survival by Sex (Barplot)**: Females: ~74% survival; males: ~19% survival.
- **Survival by Pclass (Barplot)**: 1st class: ~63% survival; 2nd: ~47%; 3rd: ~24%.

3. Key Findings

- **Patterns and Trends**:
    - **Gender**: Females had significantly higher survival rates due to the “women and children first” policy.
    - **Class and Wealth**: 1st class passengers and those paying higher fares had better survival odds, reflecting socio-economic privilege.
    - **Age**: Children had higher survival rates, while adults (20-40) dominated the passenger distribution.
    - **Family Size**: Small families (1-2 members) had better survival than lone travelers or large families.
    - **Embarkation**: Cherbourg passengers had higher survival, likely due to more 1st class passengers.

- **Anomalies**:
    - Missing data in age and cabin requires imputation or exclusion for modeling.
    - Extreme fares (e.g., 512) and rare ages (e.g., infants, 80-year-olds) are outliers.
- **Relationships**:
    - Negative correlation between pclass and survived (-0.34).
    - Positive correlation between fare and survived (0.26).
- **Overall Survival**: ~38% survival rate, with clear biases toward females, higher classes, and wealthier passengers.

4. Conclusion

This EDA reveals that survival on the Titanic was heavily influenced by gender, socio-economic status, and age. The analysis highlights the need for data cleaning (e.g., handling missing age values) and provides a foundation for predictive modeling. The visualizations and statistical summaries effectively uncover patterns and anomalies, building skills in data exploration using Python.