Analysis Report: Titanic Dataset

By Kushagra Agrawal ECE 2027 23116051

1. Dataset Overview

- **Size**: 891 passengers, 15 features.
- **Target Variable**: Likely `survived` (binary: 0 = died, 1 = survived) or `alive` (no/yes).
- **Key Features**:
  - Demographics: `sex`, `age`, `who` (man/woman/child), `adult_male`.
  - Socioeconomic: `pclass`, `class`, `fare`.
  - Travel details: `embarked`, `embark_town`, `alone`, `sibsp` (siblings/spouses), `parch` (parents/children).
  - Survival metadata: `survived`, `alive`, `deck`.

2. Data Quality Issues

Missing Values:

- **Age**: 177 missing values (20% of data). Addressed by median imputation (29.36 years).
- **Deck**: 688 missing (77% of data). Likely excluded due to high incompleteness.
- **Embarked/Embark Town**: 2 missing. Filled with mode (`S`/Southampton).
- **Action**: `deck` should be dropped; other imputations are reasonable.

Duplicates:

- 107 duplicate rows (12% of data). Removed to avoid bias, though some duplicates might reflect families/groups.

Outliers:

- **Age**: 27 outliers (e.g., very young children or older adults).

- **Fare**: 102 outliers (high-cost tickets). Likely reflects luxury-class passengers.
- **Action**: Retained outliers as they may represent valid edge cases (e.g., wealthy passengers).

3. Key Distributions

Demographics:

- **Sex**: 62% male, 38% female.
- **Age**: Mean = 29.4 years (SD = 13.2), right-skewed (skew = 0.21). Majority aged 21–36.
- **Who**: 57% men, 32% women, 11% children.

Socioeconomic Status:

- **Class**: 52% Third Class, 27% First Class, 21% Second Class.
- **Fare**: Mean = £26.6 (SD = £22.9), highly skewed (skew = 1.1). 75% paid ≤ £34.2, max £73.4.

Survival:

- **Overall Survival Rate**: 41.3% survived (`survived` = 1).
- **Alive vs. Dead**: 59% died (`alive` = no), 41% survived (`alive` = yes).

Travel Details:

- **Embarkation**: 72% Southampton, 20% Cherbourg, 8% Queenstown.
- **Alone**: 60% traveled alone (`alone` = True).

4. Survival Analysis

By Sex:

- **Hypothesis**: Women and children survived at higher rates ("women and children first" policy).

- **Data**: `who` column shows 32% women and 11% children. Survival rates likely correlate with `sex` and `who`.

By Class:

- **Hypothesis**: Higher survival in First Class.
- **Data**: 27% of passengers were First Class, but survival statistics (mean `survived` = 0.41) suggest class impacted outcomes.

By Fare:

- **Outliers**: High-fare passengers (outliers) likely in First Class. Survival rates may correlate with fare.

By Age:

- Children (mean age = 29.4) may have higher survival rates. Age distribution skews young, but missing data complicates analysis.

---

5. Statistical Insights

- **Skewness/Kurtosis**:
  - `sibsp` (skew = 3.0) and `parch` (skew = 2.6): Most passengers had 0 siblings/spouses or parents/children.
  - `fare` (skew = 1.1): Confirms socioeconomic disparity.
- **Correlations** (Inferred):
  - Negative correlation between `pclass` and `survived` (lower class = lower survival).
  - Positive correlation between `fare` and `survived`.

---

6. Visualization Insights

(Assuming plots were generated:)

- **Survival by Class**: First Class passengers had higher survival rates.
- **Survival by Sex**: Females survived more than males.

- **Age Distribution**: Bimodal peaks for children (0–10) and adults (20–40).
- **Fare vs. Survival**: Higher fares linked to survival.

7. Limitations

1. **Missing Data**: Age and deck missingness may bias results.
2. **Duplicates**: Removal may exclude family groups.
3. **Outliers**: High fares and ages retained but may skew statistics.
4. **Categorical Features**: `embarked`, `class`, and `who` need encoding for modeling.

8. Recommendations

1. **Feature Engineering**:
   - Create `family_size` from `sibsp + parch`.
   - Bin `age` into groups (child/adult/senior).
   - Encode `embarked`, `class`, and `who` for ML.
2. **Modeling**:
   - Use logistic regression or decision trees to predict survival.
   - Key predictors: `pclass`, `sex`, `fare`, `age`.
3. **Further Analysis**:
   - Investigate interaction effects (e.g., `class` × `sex`).
   - Explore survival rates by `embark_town` and `alone`.

9. Conclusion

The dataset reveals clear socioeconomic and demographic disparities in survival:

- **Women, children, and First Class passengers** had higher survival rates.
- **High fare prices** and **embarkation from Cherbourg** (linked to First Class) correlate with survival.
- **Age** and **family size** may further refine predictions.

Next Steps: Build a predictive model and validate hypotheses with statistical tests (e.g., chi-square for categorical features).