TITANIC Dataset Cleaning+Analysis

—By `KARTIKEY PAL 23116041`

Introduction

The Titanic dataset provides insights into passenger demographics, travel details, and survival rates. This report analyzes the dataset to identify patterns and trends in survival probabilities based on different features.

Dataset Description

- **Total Records**: 891 passengers
- **Target Feature**: `survived` (0 = No, 1 = Yes)
- **Feature Categories**:
    - **Demographic Information**: `sex`, `age`, `who`, `adult_male`
    - **Social & Economic Factors**: `pclass`, `fare`, `class`
    - **Travel Details**: `embarked`, `alone`, `sibsp`, `parch`
    - **Survival Indicators**: `survived`, `alive`, `deck`

Data Quality & Cleaning

Handling Missing Values:

- `age`: 20% missing; filled with median (29.36 years).
- `deck`: 77% missing; dropped due to excessive missingness.
- `embarked` & `embark_town`: 2 missing values; imputed using mode (`S`).

Duplicate Entries:

- 107 duplicate rows detected and removed.

Outliers:

- `age` and `fare` have extreme values, but these are retained as they represent valid cases.

Data Distribution Analysis

Demographic Breakdown:

- **Gender**: 62% male, 38% female
- **Age Distribution**: Mean = 29.4 years, most passengers aged 21–36
- **Passenger Categories**: 57% men, 32% women, 11% children

Economic & Class Information:

- **Class Proportions**: 52% Third Class, 27% First Class, 21% Second Class
- **Fare Statistics**: Mean = £26.6, highly skewed; 75% paid ≤ £34.2

Survival Rate:

- **Overall Survival**: 41.3%
- **Survived vs. Deceased**: 59% perished, 41% survived

Embarkation Details:

- **Embarked From**: 72% Southampton, 20% Cherbourg, 8% Queenstown
- **Solo Travelers**: 60% of passengers traveled alone

Factors Affecting Survival

Survival by Gender:

- Women and children had higher survival rates than men.
- `who` column confirms survival disparity by gender.

Survival by Class:

- First-Class passengers had a significantly higher survival rate.

Fare Influence:

- Higher fares correlate with better survival chances, as wealthier passengers had priority access to lifeboats.

Impact of Age:

- Younger passengers had better survival rates, but missing data limits precision.

Statistical Insights

Skewness & Data Distribution:

- `sibsp` (skew = 3.0) and `parch` (skew = 2.6) suggest most passengers traveled alone.
- `fare` is right-skewed, indicating high ticket price variability.

Correlation Trends:

- **Negative correlation** between `pclass` and `survived`
- **Positive correlation** between `fare` and `survived`

Key Visualizations & Trends

(Assumed visualizations used in analysis)

- **Class vs. Survival**: Higher-class passengers had better survival rates.
- **Gender vs. Survival**: Women outlived men significantly.
- **Age Distribution**: Peaks at childhood (0–10) and adulthood (20–40).
- **Fare & Survival**: Expensive tickets linked to survival.

Challenges & Limitations

1. **Data Gaps**: High missing values for `deck`, some for `age`.
2. **Duplicate Removal**: May remove passengers from the same group.
3. **Outliers**: Extreme values for `fare` and `age` retained but could affect modeling.
4. **Categorical Encoding**: Features like `embarked`, `who`, and `class` need transformation for ML models.

Suggested Enhancements

1. **Feature Engineering:**
    - Create `family_size = sibsp + parch`
    - Bin `age` into categories (child, adult, senior)
    - Encode categorical features for predictive modeling
2. **Predictive Modeling:**
    - Apply logistic regression or decision trees for survival prediction
    - Use `pclass`, `sex`, `fare`, and `age` as key predictors
3. **Further Research:**
    - Investigate impact of `embark_town`
    - Examine effects of traveling alone vs. with family

Final Thoughts

The Titanic dataset highlights clear survival patterns:

- **Women, children, and First-Class passengers had the highest survival rates.**
- **High fares and embarkation from Cherbourg correlate with survival.**
- **Age and family size are important secondary factors.**

Future Direction: Develop a predictive model and validate findings with hypothesis testing (e.g., chi-square for categorical relationships).