Report of Findings

Observations:

- Dataset Size: 891 passengers with 12 features
- Target Variable: Survived (0 = died, 1 = survived)
- Data Types: Mix of numeric (7) and categorical (5) variables
- Survival Rate: 38.4% overall (342 survived out of 891)

Data Quality Assessment

Missing Values Analysis:

- Cabin: 687 missing (77.1%) - Highest missingness, likely indicates lower-class passengers
- Age: 177 missing (19.9%) - Significant but manageable
- Embarked: 2 missing (0.2%) - Minimal impact
- No Duplicate Records: Data integrity is good

Visual Analysis and Observations

1. Survival Distribution (Histogram)

Observations:

- Clear imbalance: ~549 died vs ~342 survived
- Survival rate of 38.4% reflects the tragic nature of the disaster
- Binary distribution shows this as a classification problem

2. Passenger Class Distribution (Histogram)

Observations:

- 3rd Class dominates: ~500 passengers (56%)
- 1st Class: ~200 passengers (22%)
- 2nd Class: ~180 passengers (20%)
- Reflects socioeconomic structure of early 1900s ocean travel

3. Age Distribution (Histogram)

Observations:

- Right-skewed distribution with median around 28 years
- Peak at 20-30 years: Most passengers were young adults
- Long tail: Some elderly passengers up to 80 years
- Few infants/children: Small peak near 0-5 years

4. SibSp (Siblings/Spouses) Distribution

Observations:

- Heavily skewed: Most passengers traveled alone or with 1 family member
- Peak at 0: ~600 passengers had no siblings/spouses aboard
- Rapid decline: Very few traveled with large families
- Maximum of 8 siblings/spouses (rare cases)

5. Parch (Parents/Children) Distribution

Observations:

- Even more skewed than SibSp: ~680 passengers traveled without parents/children
- Most common: 0 or 1 parent/child
- Family structure: Suggests many were solo travelers or young couples

6. Fare Distribution (Histogram)

Observations:

- Heavily right-skewed: Most passengers paid low fares
- Peak at low end: Majority paid £0-50
- Long tail: Some extremely expensive tickets (up to £500+)
- Clear class distinction: Reflects passenger class differences

7. Gender Distribution (Bar Chart)

Observations:

- Male majority: ~577 males (64.8%) vs ~314 females (35.2%)
- Gender imbalance: Typical of early 1900s sea travel
- Critical for survival analysis: Gender was a key factor in "women and children first" protocol

8. Embarked Port Distribution (Bar Chart)

Observations:

- Southampton (S) dominates: ~644 passengers (72.4%)
- Cherbourg (C): ~168 passengers (18.9%)
- Queenstown (Q): ~77 passengers (8.6%)
- Geographic pattern: Reflects the ship's route from Southampton to New York

9. Correlation Heatmap Analysis

Key Correlations Observed:

- Pclass vs Fare: Strong negative correlation (-0.55) - higher class = higher fare
- Age vs Pclass: Weak negative correlation - older passengers in higher classes

- SibSp vs Parch: Moderate positive correlation (0.41) - family groups travel together
- Survived vs Pclass: Negative correlation (-0.34) - higher class = better survival
- Survived vs Fare: Positive correlation (0.26) - higher fare = better survival

10. Scatter Matrix Analysis

Observations from Top Variance Features (PassengerId, Fare, Age, SibSp, Pclass):

- Fare vs Pclass: Clear inverse relationship visible
- Age distribution: Relatively normal across different classes
- Family size patterns: Most passengers traveled in small groups
- PassengerId: Random distribution (just an identifier)

Summary of Key Findings

Critical Survival Factors Identified:

1. Social Class Impact
   - Strong correlation between passenger class and survival
   - Higher fare passengers had better survival chances
   - Economic status was a major survival predictor
2. Demographic Patterns
   - Gender imbalance (65% male) suggests different survival rates likely
   - Age distribution shows predominantly young adult passengers
   - Family structure varies significantly across passengers
3. Geographic and Social Context
   - Southampton was the primary departure point
   - Clear socioeconomic stratification visible in fare distribution
   - Family travel patterns suggest different survival strategies

Data Quality Insights:

1. Missing Data Strategy Needed
   - Cabin data missing for 77% - requires careful handling
   - Age imputation needed for 20% of passengers
   - Embarked has minimal missing data

2. Feature Engineering Opportunities
    - Combine SibSp and Parch for family size
    - Create age groups for better analysis
    - Extract titles from names for social status
    - Create fare buckets for class analysis

🔍 Statistical Implications:

1. Strong Correlations Found
    - Class and fare are inversely related (as expected)
    - Clear socioeconomic patterns in the data
    - Family structure shows logical patterns
2. Survival Prediction Potential
    - Multiple features correlate with survival
    - Class, fare, and demographics show promise
    - Feature engineering could improve prediction accuracy