Titanic Dataset Analysis Insights

1. **Data Understanding and Preparation**
    a. **Dataset Overview**
        i. The dataset contains 891 rows and 12 columns (11 features + 1 target variable "Survived").
    b. **Missing Values**
        i. Age: 19.87% missing values. These rows were dropped.
        ii. Cabin: 77.10% missing values. The missing values were imputed using the mode (most frequent value).
        iii. Embarked: 0.22% missing values. These were also imputed using the mode.
    c. **Data Types**
        i. The Age column was converted from float64 to int64 since age is typically represented as an integer.
    d. **Outliers**
        i. No significant outliers or invalid values were detected in the dataset.
2. **Data Visualization and Bivariate Analysis**
    a. **Pclass vs Survived**
        i. Passengers with 1st class tickets had the highest survival rate.
        ii. Passengers with 3rd class tickets had the lowest survival rate.
        iii. 2nd class passengers had mixed survival rates.
    b. **Sex vs Survived**
        i. Females had a significantly higher survival rate compared to males.
        ii. This is likely due to the "women and children first" protocol during the evacuation.
    c. **Age vs Survived**
        i. Age was binned into categories: 020, 2140, 4160, and 6180.
        ii. Passengers aged 2140 had the highest death count.
        iii. Passengers aged 6180 had the lowest death count.
        iv. The high death count in the 2140 age group is likely due to the majority of passengers in this group being 3rd class males, who had lower survival rates.
    d. **SibSp (Siblings/Spouses) vs Survived**
        i. Passengers with 0, 2, 3, 4, or 5 siblings/spouses had higher death rates.
        ii. Passengers with 1 sibling/spouse had a higher survival rate.
    e. **Parch (Parents/Children) vs Survived**
        i. Passengers with 0, 4, 5, or 6 parents/children had higher death rates.
        ii. Passengers with 1, 2, or 3 parents/children had higher survival rates.
    f. **Embarked (Port of Embarkation) vs Survived**
        i. Passengers who embarked from Southampton (S) and Queensland (Q) had higher death rates.
        ii. Passengers who embarked from Cherbourg (C) had higher survival rates

iii. This is likely because Southampton had a higher number of 3rd class passengers and males, both of which had lower survival rates.
iv. Cherbourg had a higher proportion of 1st class passengers, who had higher survival rates.

3. Feature Engineering

a. Age Binning

i. The Age column was binned into categories (020, 2140, 4160, 6180) to simplify analysis.

b. Categorical to Ordinal Conversion

i. Sex: Converted to binary (0 for male, 1 for female).
ii. Embarked: Converted to ordinal (0 for Cherbourg, 1 for Queensland, 2 for Southampton).

c. Irrelevant Columns Dropped

i. Columns like Fare, Ticket, PassengerId, Cabin, and Name were dropped as they were deemed irrelevant for prediction.

4. Feature Importance

a. ChiSquare Test

i. The most important features based on the ChiSquare test were Sex, Pclass, Parch, and Embarked.
ii. Sex had the highest ChiSquare score, indicating it is the most significant predictor of survival.

b. Correlation Matrix

i. Sex was positively correlated with survival, indicating that females had a higher chance of surviving.
ii. Pclass was negatively correlated with survival, indicating that lower class passengers had a lower chance of surviving.
iii. Other variables like Parch, SibSp, Age_Bins, and Embarked also showed some correlation with survival.

5. Key Observations and Conclusions

a. Survival Factors

i. **Sex**: Females had a much higher survival rate than males.
ii. **Pclass**: 1st class passengers had the highest survival rate, while 3rd class passengers had the lowest.
iii. **Age**: Younger passengers (020) and older passengers (6180) had higher survival rates compared to middle aged passengers (2140).
iv. **Embarked**: Passengers from Cherbourg had higher survival rates compared to those from Southampton and Queensland.

b. Feature Importance

i. Sex and Pclass were the most important features in predicting survival.

c. Data Preparation

i. The dataset was cleaned by handling missing values, converting categorical variables to ordinal, and dropping irrelevant columns.

Summary of Key Insights

- Survival was highly influenced by Sex, Pclass, and Age.
- Females, 1st class passengers, and younger/older passengers had higher survival rates.
- Passengers from Cherbourg had better survival chances compared to those from Southampton and Queensland.
- Feature engineering and data preparation were crucial in improving the model's predictive power.