Data Analytics Internship

Day5   Exploratory Data Analysis (EDA)

Dataset Used: Titanic_Datatset.csv

1. Basic Exploration using NumPy and Pandas

- Used functions: **`.head()`**, **`.info()`**, **`.describe()`**, **`.isnull().sum()`**
- Outcome: Helped understand the basic structure of the dataset including data types, number of rows/columns, summary statistics, and missing values.
- Inference: Detected missing values in '**Age**', '**Fare**', and '**Embarked**' columns. Noted that data had a mix of categorical and numerical variables.
- Additional Computations:
  - Calculated percentage of missing values to prioritize cleaning tasks.
  - Performed survival count and survival rate to understand class imbalance.
  - Distribution of Categorical Variables using **`.value_counts()`** provided insights into the passenger distribution by gender, class, embarkation point, etc.

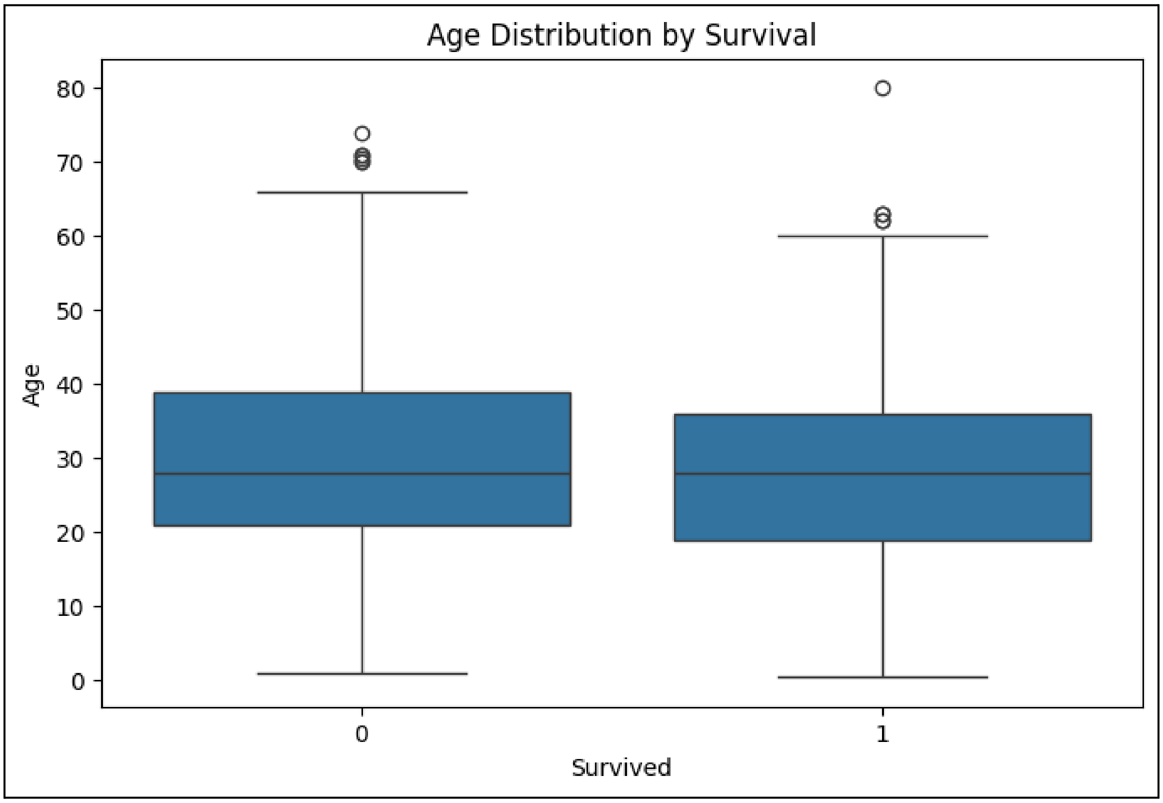
2. Data Cleaning

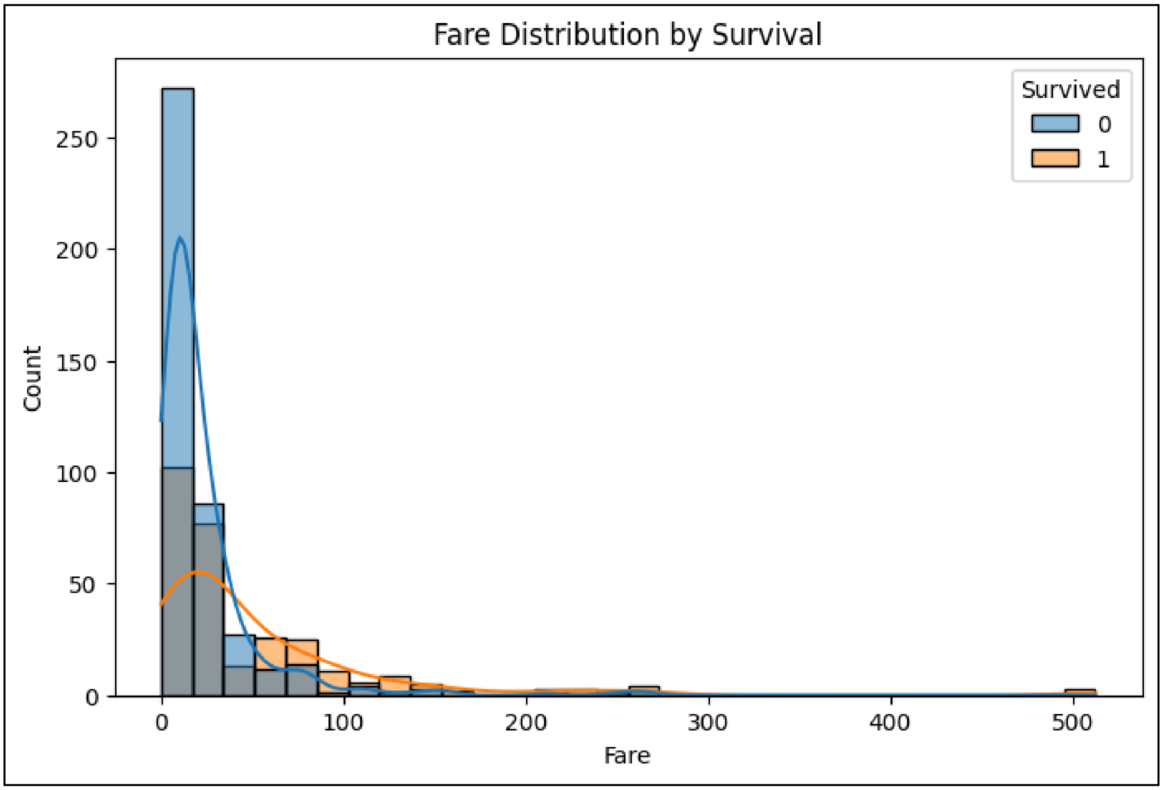
- Filled missing '**Embarked**' values using mode.
- Encoded categorical variables such as '**Sex**' and **'Embarked'** using **`.map()`.**
- Dropped rows with missing '**Age**' and '**Fare**' using **`.dropna()`**.
- Outcome: Cleaned dataset ready for analysis and visualization.
- Inference: Encoding allowed easier plotting and modeling; cleaning ensured no skew from missing data.

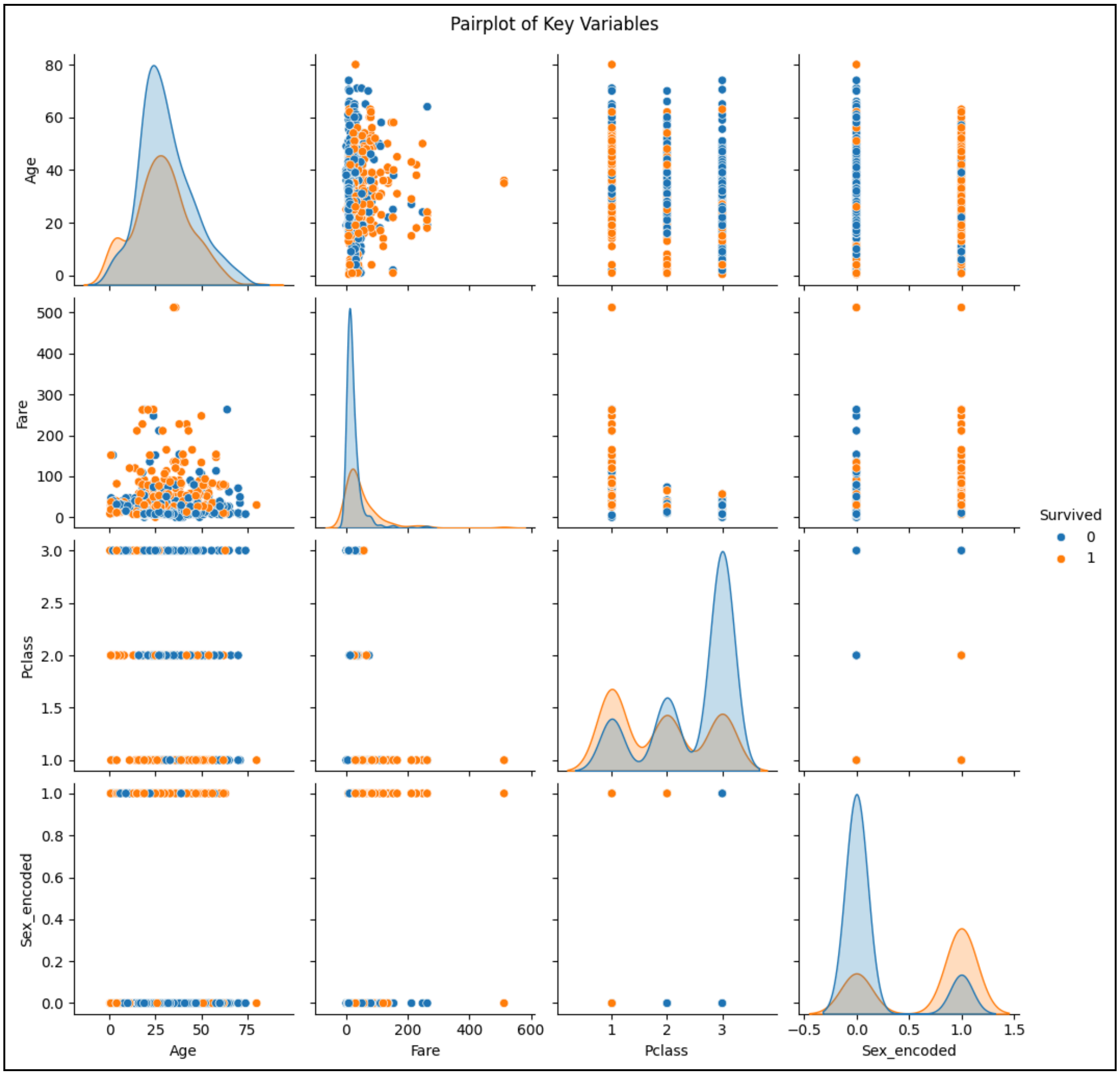
3. Graphical Exploration

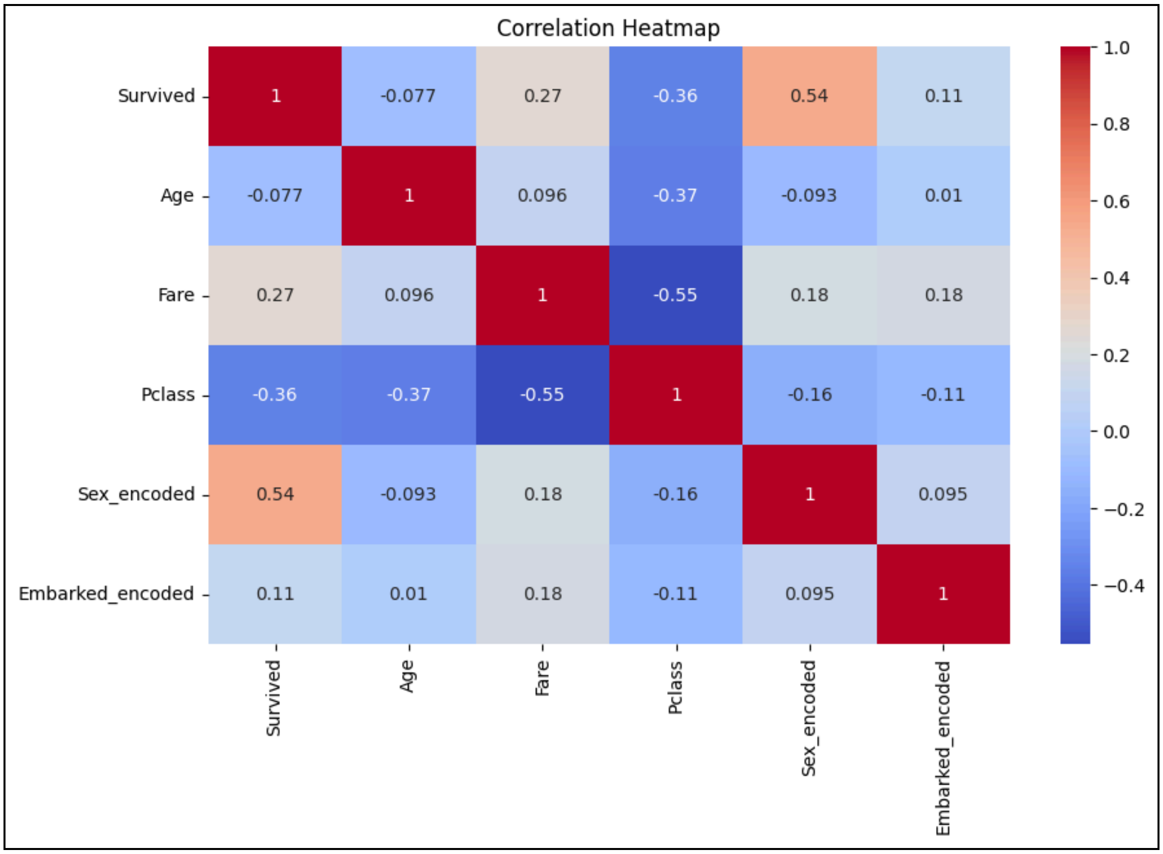
- **Boxplot (Age vs Survival)**, **Histogram (Fare with Survival hue)**, **Pairplot,** and **Heatmap using Seaborn**:
  - Outcome: Visualized survival relationships with other features.
  - Inference: Females, higher class passengers, and younger individuals had higher survival chances.
- **Countplot of Passenger Class vs Survival**:
  - Outcome: Clear comparison between survival rate across passenger classes.

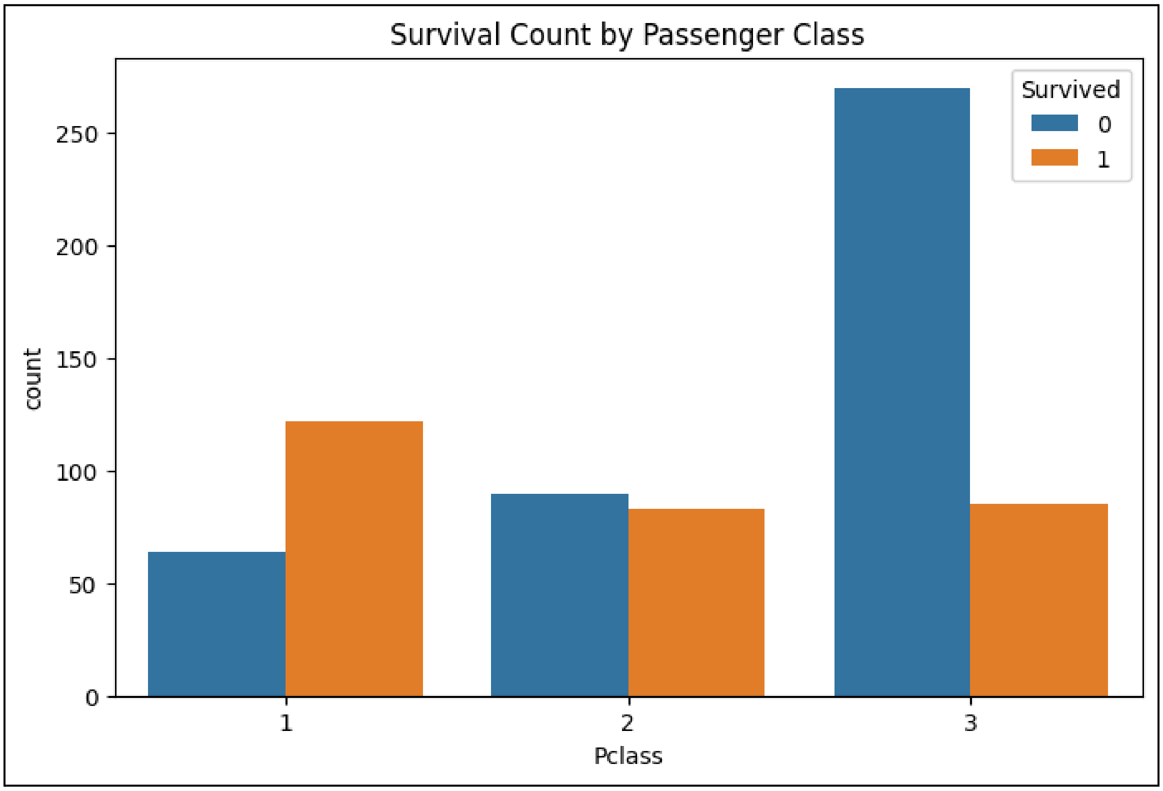
- Inference: Passengers from 1st class had significantly higher survival rates.





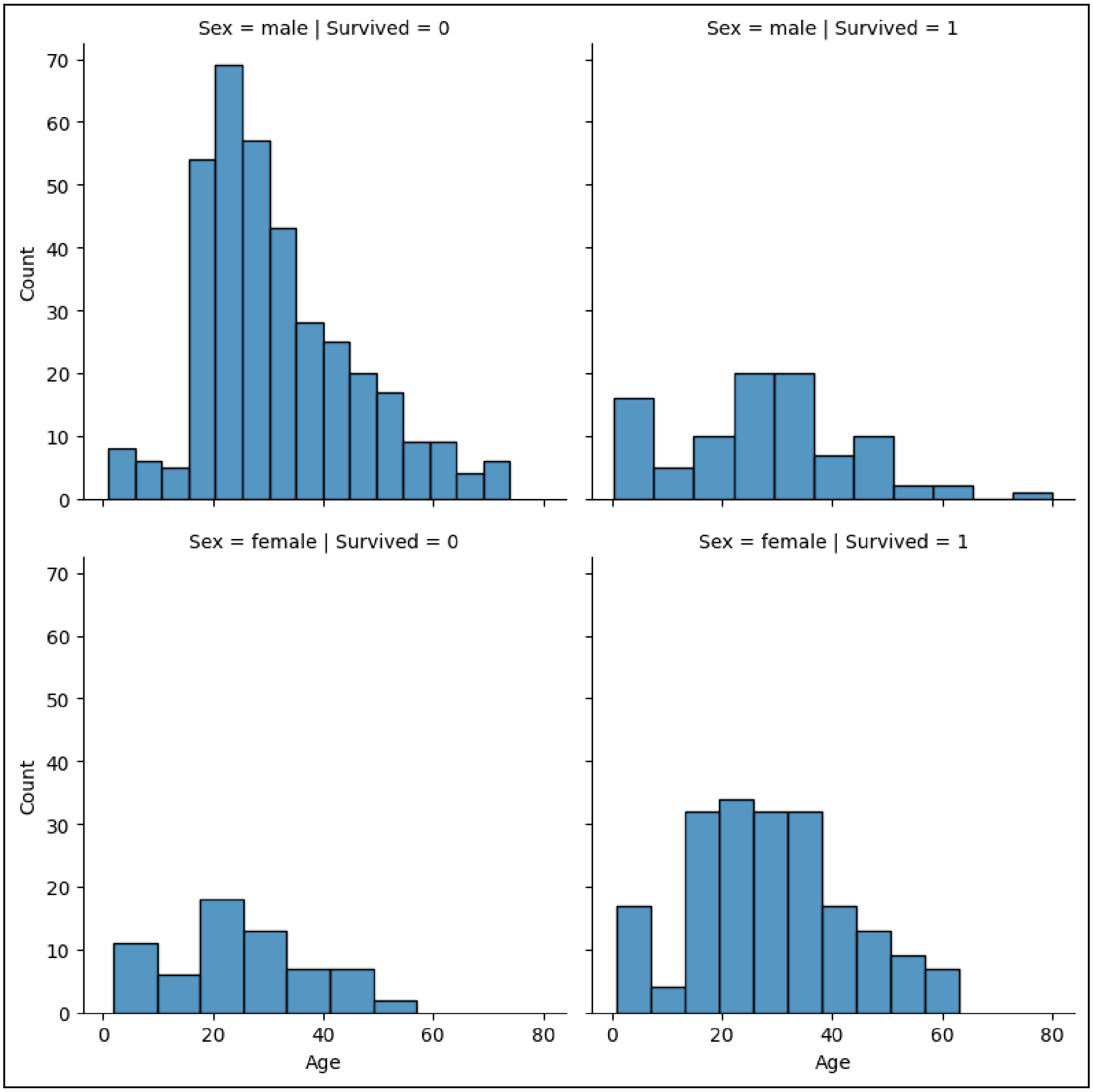

Pairplot of Key Variables
Age
0
20
40
60
80
Fare
0
100
200
300
400
500
Pclass
1.0
1.5
2.0
2.5
3.0
Sex_encoded
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
Age
0
200
400
600
Fare
1
2
3
Pclass
−0.5
0.0
0.5
1.0
1.5
Sex_encoded
Survived
0
1

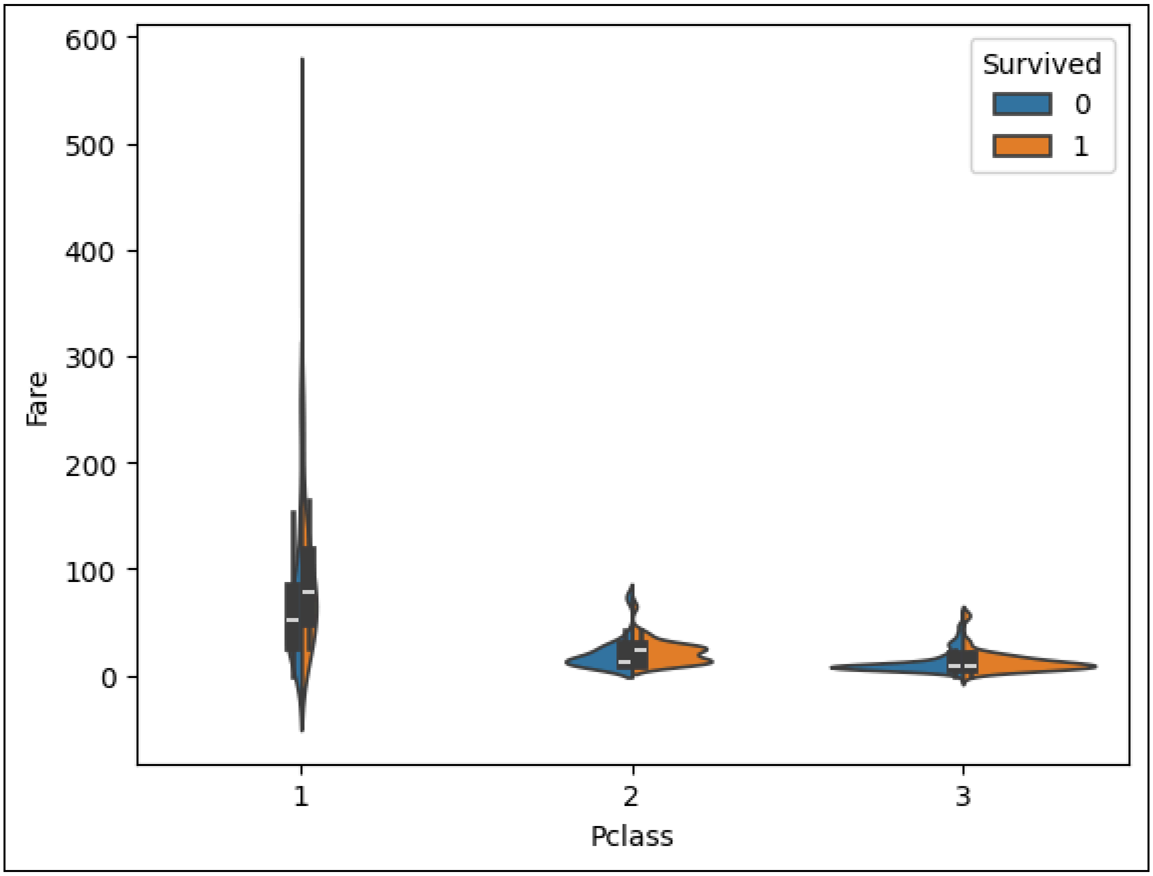

Correlation Heatmap
Survived
Age
Fare
Pclass
Sex_encoded
Embarked_encoded
1
-0.077
0.27
-0.36
0.54
0.11
-0.077
1
0.096
-0.37
-0.093
0.01
0.27
0.096
1
-0.55
0.18
0.18
-0.36
-0.37
-0.55
1
-0.16
-0.11
0.54
-0.093
0.18
-0.16
1
0.095
0.11
0.01
0.18
-0.11
0.095
1
Survived
Age
Fare
Pclass
Sex_encoded
Embarked_encoded
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Survival Count by Passenger Class
Survived
0
1
count
250
200
150
100
50
0
1
2
3
Pclass

4. Extra Insights & Feature Engineering

- Created new features:
  - **Family Size** (SibSp + Parch + 1)
  - **Is Alone** (Binary feature if FamilySize == 1)
  - **Title Extraction from Name**
  - **Age Group Categorization**
- **ChiSquare Tests:** Assessed relationship between categorical features (e.g., Gender, Pclass) and Survival.
- **Violin Plot (Fare vs Class vs Survival)** and **FacetGrid (Age & Sex vs Survival)** for deeper insights.
  - Outcome: Derived new columns and revealed more meaningful survival patterns.
  - Inference: Certain titles (like 'Mrs', 'Miss') and family configurations were more likely to survive.


Sex = male | Survived = 0
Sex = male | Survived = 1
Sex = female | Survived = 0
Sex = female | Survived = 1
Count
Count
0
10
20
30
40
50
60
70
0
20
40
60
80
Age
Age



5. Final Summary & Observations

- Data cleaning and preprocessing were critical to prepare the dataset.
- Visualizations provided both univariate and multivariate insights.
- Engineered features enhanced understanding of survival patterns.
- Gender, class, family structure, and embark location strongly influenced survival.
- Navigation techniques, clear layout, and structured visuals were used for better storytelling.