Date:11-082025

Titanic Dataset – Exploratory Data Analysis (EDA) Report

Executive Summary:

This report presents the key findings from an exploratory data analysis (EDA) of the Titanic dataset. It covers the entire step-by-step process, from data loading to correlation analysis, and highlights trends, imbalances, correlations, and notable patterns that influenced passenger survival.

Step-by-Step EDA Process

Step 1: Data Loading - Loaded the Titanic dataset into a Pandas DataFrame for exploration.

Step 2: Initial Inspection - Checked the shape, data types, and previewed first few rows.

Step 3: Missing Value Analysis - Identified missing values in 'Age', 'Cabin', and 'Embarked'.

Step 4: Handling Missing Values - Filled 'Age' with median, 'Embarked' with mode, and dropped/flagged 'Cabin'.

Step 5: Feature Understanding - Reviewed each column's meaning and relevance.

Step 6: Univariate Analysis - Plotted distributions of numerical and categorical columns.

Step 7: Observations from Categorical Features - Sex imbalance, embarked port imbalance, ticket uniqueness.

Step 8: Observations from Numerical Features - Age skew toward younger adults, right-skewed fare distribution.

Step 9: Bivariate Analysis - Examined survival rates by sex, age, class, fare.

Step 10: Multivariate Analysis - Used scatter plots and heatmaps to explore feature interactions.

Step 11: Correlation Analysis - Found strong negative correlation between Pclass and Fare, positive between SibSp and Parch.

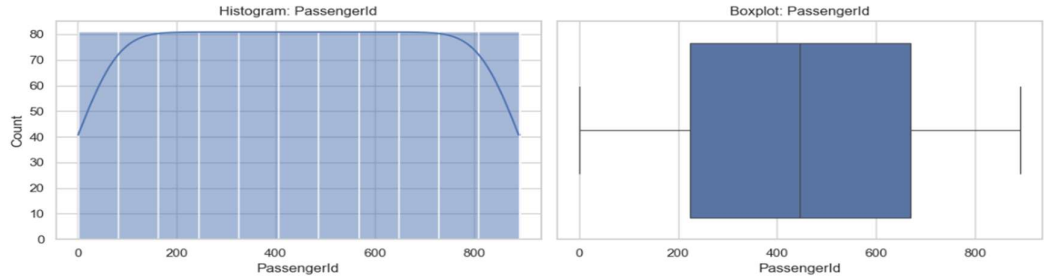
Step 12: Key Insights - Women, children, and higher-class passengers had better survival chances.

Key Findings & Observations:

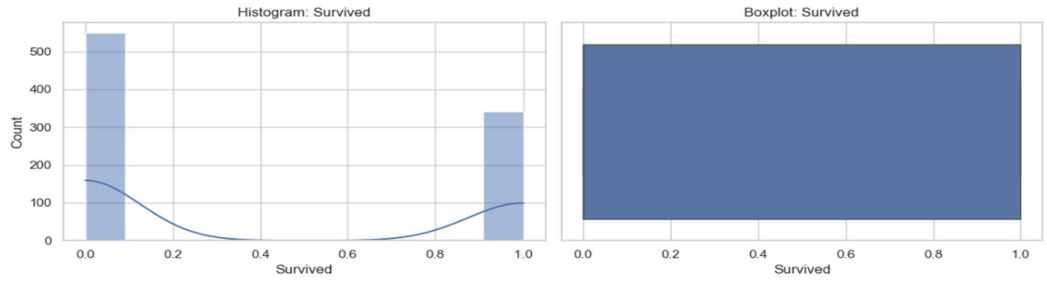
- ❖ Passenger IDs are unique identifiers and have no influence on survival outcomes. Names are distinct for each passenger and serve mainly for reference.
- ❖ There is a sex imbalance, with significantly more male passengers than females.
- ❖ Most passengers were in their 20s–30s, with fewer children and elderly passengers.
- ❖ Fares show a right-skewed distribution; higher fares are linked to higher-class tickets.
- ❖ Tickets are mostly unique, limiting their usefulness for grouping passengers.
- ❖ Most passengers traveled alone or with just 1–2 family members.
- ❖ Most passengers boarded at port S, fewer at ports C and Q.
- ❖ Women and children had higher survival rates compared to men and older passengers.
- ❖ Younger passengers who paid higher fares had better survival chances.
- ❖ Pclass is negatively correlated with Fare, and SibSp is positively correlated with Parch. Other correlations are weak.

Visual Insights

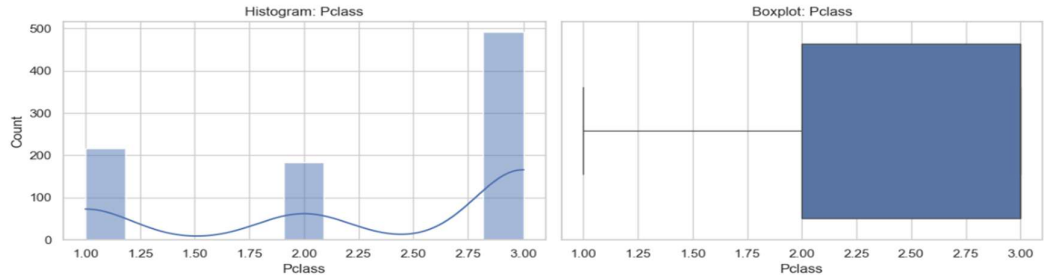
The following visualizations were created during the analysis:



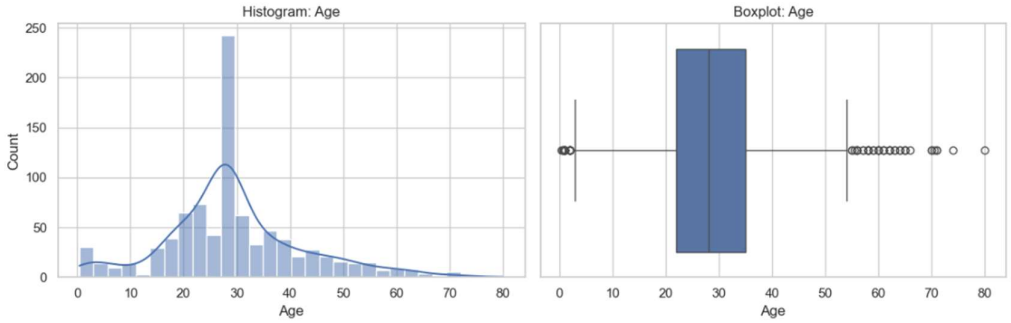
Observation (PassengerId): Each passenger has a unique ID number that simply helps us keep track of them in the dataset. It's basically a serial number, so it doesn't carry any real-world meaning beyond identification.



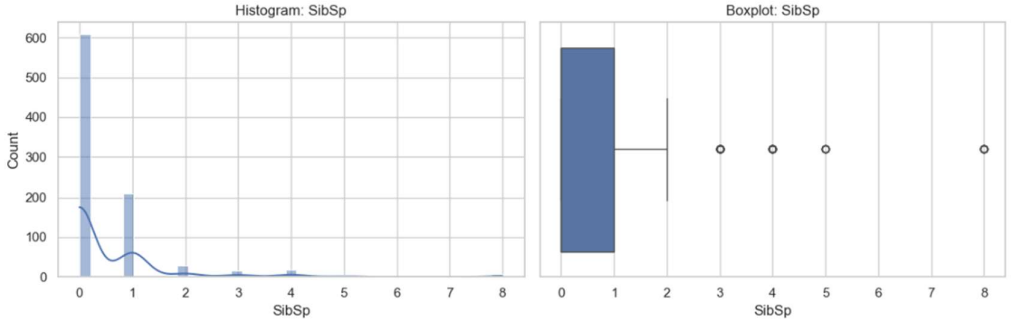
Observation (Survived): This column tells us whether a passenger survived (1) or not (0). It's not a continuous number, just a survival flag, so we'll use it later mainly for classification, not distribution analysis.



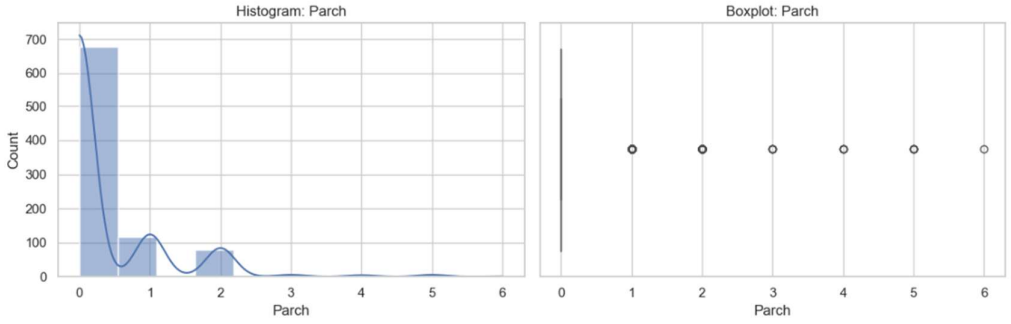
Observation (Pclass): The passenger class is like the travel tier — 1st, 2nd, or 3rd class. Most passengers seem to be in 3rd class, which could say something about affordability or demographics at the time.



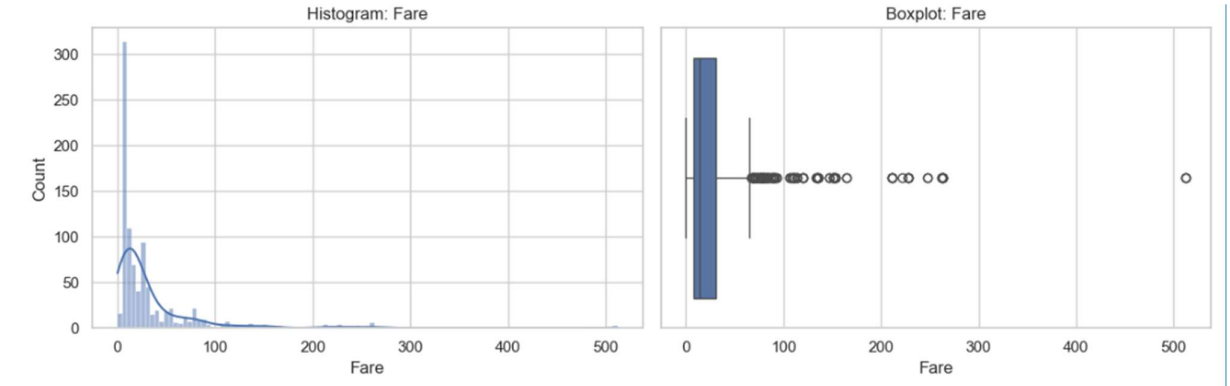
)bservation (Age): Passenger ages range from infants to elderly travelers, but the distribution shows many passengers were young adults. There are also some missing alues we'll need to handle before analysis.



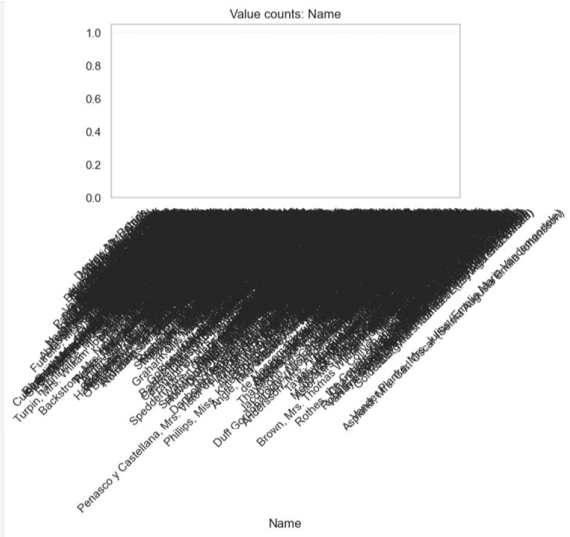
)bservation (SibSp): This shows how many siblings or spouses a passenger had aboard. Most passengers traveled alone (0), but some had one or more family members vith them.



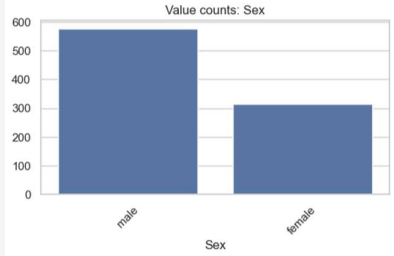
)bservation (Parch): This tells us how many parents or children a passenger had aboard. Like SibSp, most passengers had 0 here, meaning no immediate family in that ategory on board.



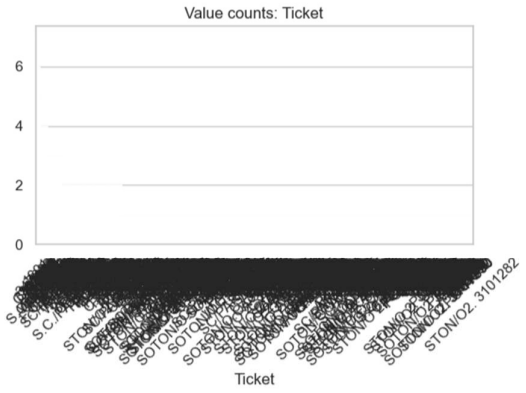
Observation (Fare): This is the ticket price, which varies widely. Most fares are on the lower side, but a few very high fares stand out — likely for luxury cabins in 1st class.



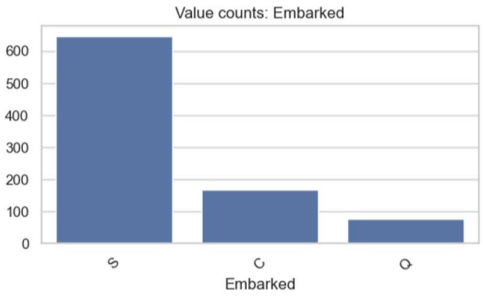
Observation (Name) : Each passenger's name is unique and mainly serves as an identifier in the dataset. It doesn't provide a measurable pattern but can be used for reference or cross-checking other details.



Observation (Sex) : There's a clear imbalance between male and female passengers, with a significantly higher number of males on board compared to females.

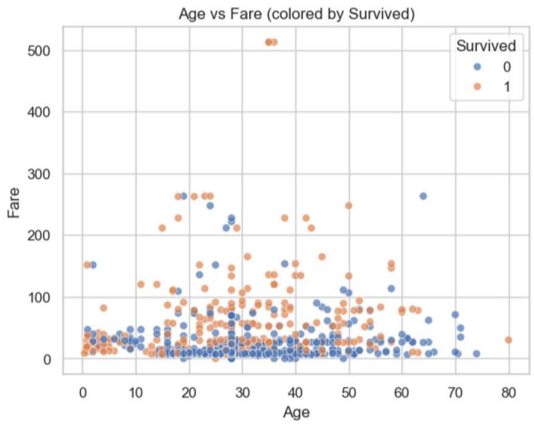


Observation (Ticket) : Ticket numbers are highly varied, with very few repeats. This means most passengers had their own unique ticket, so it's not very useful for identifying groups unless tickets are shared.



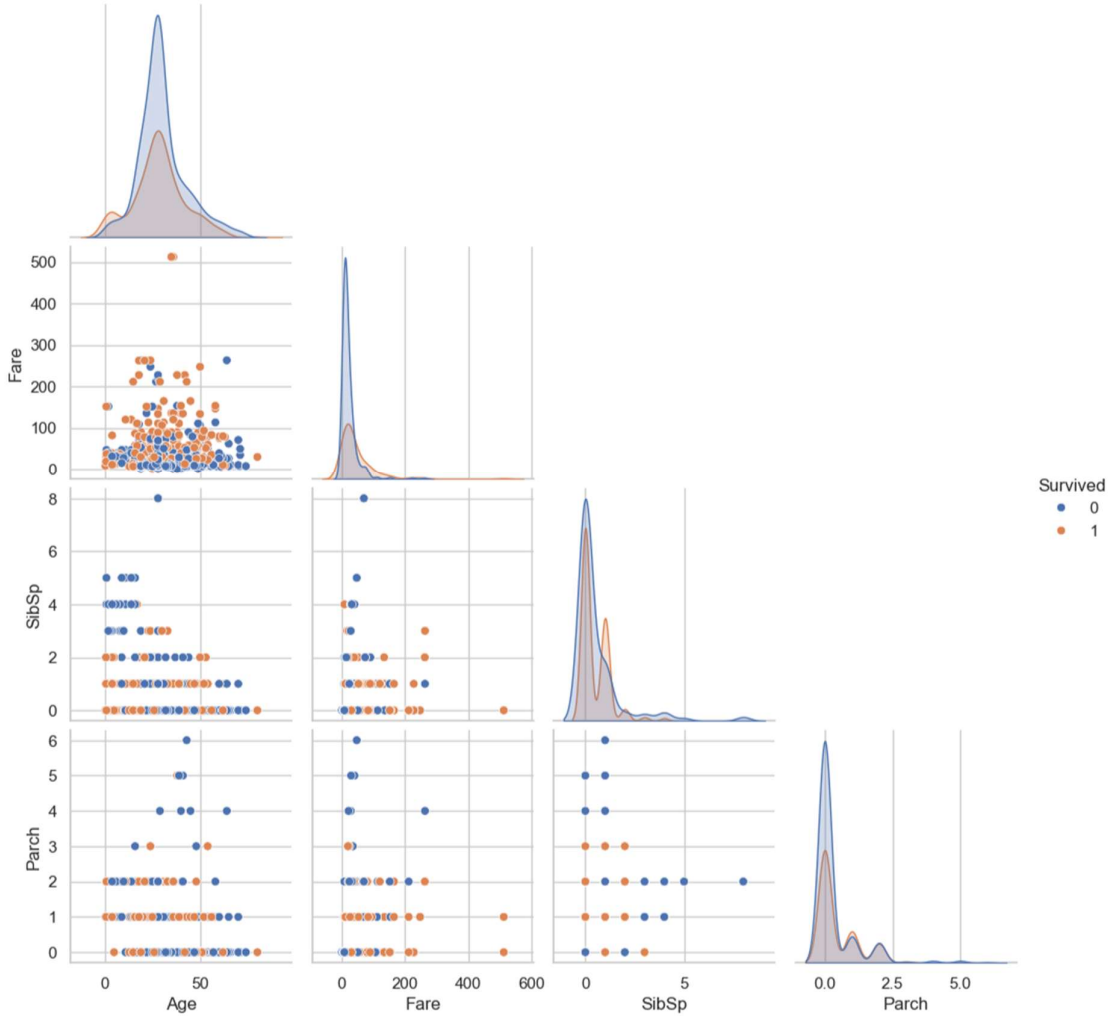
Observation (Embarked) : The majority of passengers boarded from port **S**, with much fewer from ports **C** and **Q**, showing a strong boarding location imbalance.

[27]:
```
sns.scatterplot(data=df, x='Age', y='Fare', hue='Survived', alpha=0.7)
plt.title('Age vs Fare (colored by Survived)')
plt.show()
display(Markdown("**Observation:Passengers who paid higher fares, especially younger adults, seem to have had better survival chances compared to thos
```

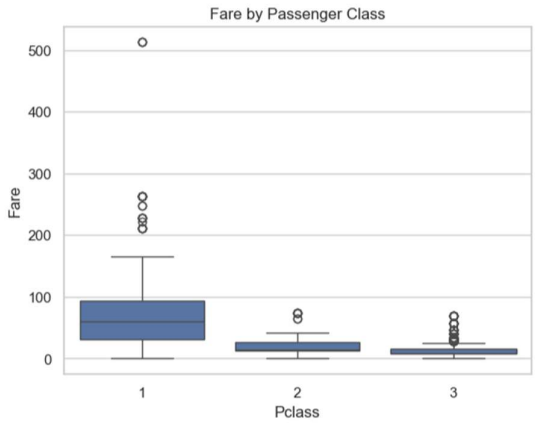


Observation:Passengers who paid higher fares, especially younger adults, seem to have had better survival chances compared to those who paid lower fares ...

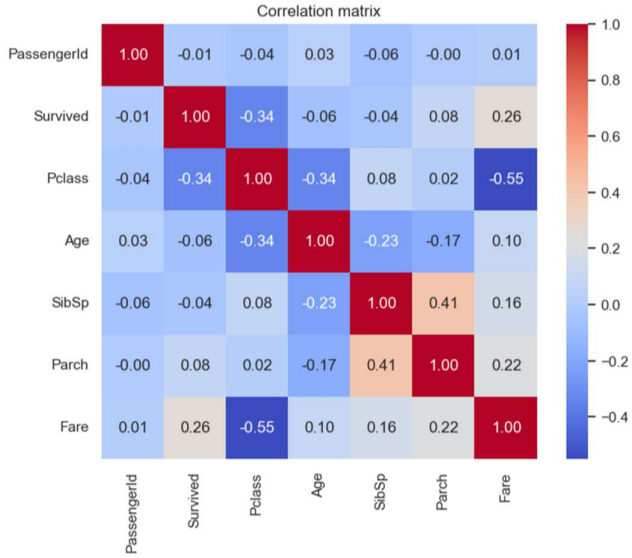
[28]:
```
sns.pairplot(df[['Age','Fare','SibSp','Parch','Survived']].dropna(), hue='Survived', corner=True)
plt.show()
```



[29]:
```
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title('Fare by Passenger Class')
plt.show()
```



[32]:
```
corr = df.select_dtypes(include=['int64','float64']).corr()
plt.figure(figsize=(8,6))
sns.heatmap(corr, annot=True, fmt=".2f", cmap='coolwarm', square=True)
plt.title('Correlation matrix')
plt.show()
display(Markdown("**Observation:** People traveling in higher classes generally paid more for their tickets, and those with more siblings/spouses aboa
```



Observation: People traveling in higher classes generally paid more for their tickets, and those with more siblings/spouses aboard also tended to have more parents/children with them, while most other factors don't show strong connections.

Dumpa Sravan Kumar

Kumarsravan3361@gmail.com