DATA ANALYST INTERNSHIP TASK 5

* Shruti A. Tembhurne

• Task 5: Exploratory Data Analysis (EDA)

• Objective: Extract insights using visual and statistical exploration.

ABOUT THE DATA:

The Titanic Passenger dataset provides information about passengers who were aboard the RMS Titanic during its ill-fated maiden voyage. This dataset is often used for exploring patterns and factors associated with survival on the Titanic.

PassengerId: Unique identifier for each passenger.  
Survived: Survival status of the passenger (0 = Not Survived, 1 = Survived).  
Pclass: Passenger class (1 = First class, 2 = Second class, 3 = Third class).  
Sex: Gender of the passenger.  
Age: Age of the passenger.  
SibSp: Number of siblings/spouses aboard the Titanic.  
Parch: Number of parents/children aboard the Titanic.  
Fare: Fare paid by the passenger.  
Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

APPROACH TO EXPLORATORY DATA ANALYSIS:

1. Download the dataset from Kaggle.

2. Create a new file named Titanic\_Data Jupyter into the Notebook.

3. Import **pandas, matplotlib, seaborn, NumPy** into the Notebook

4. Import the xlsx data file using pandas.

5. Perform Basic data exploration.

• What kind of data do we have?

• How many rows/columns are there?

• Are there any missing values?

• What are the numerical distributions?

6. Fill the null values using mode, mean etc.

7. Perform Basic data cleaning on the dataset.

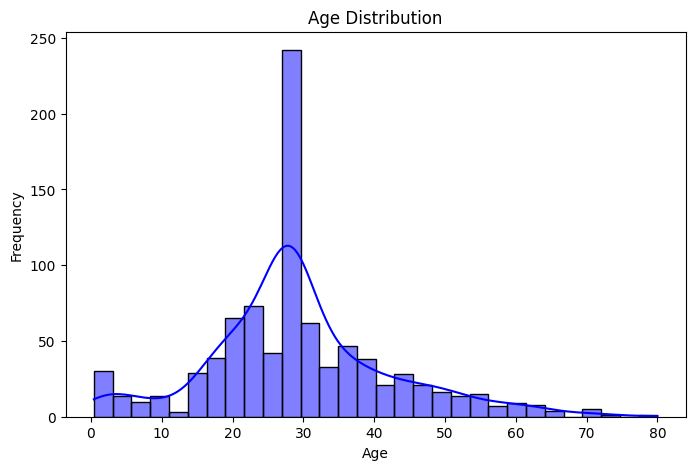
8. Plot charts using imported libraries.

9. Understand Insights of the visualization output.

10. Create a Report of the EDA.

INSIGHTS

BAR CHART OF PEOPLE SURVIVED BY AGE GROUP.



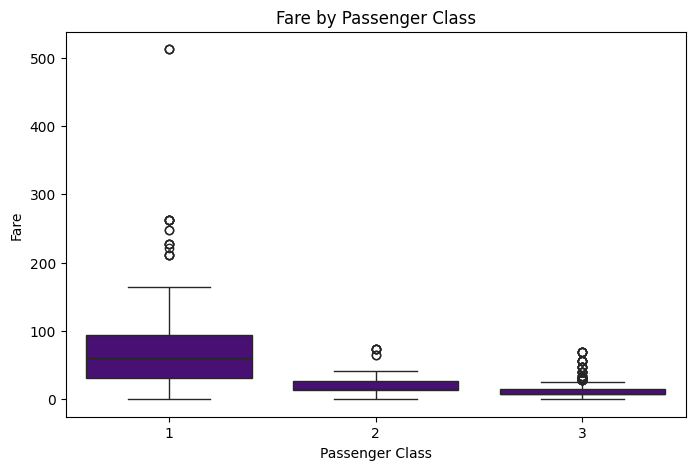
- Age distribution is right-skewed.

- Most passengers were between 20 and 40 years.

- Small number of infants and elderly aboard.

- A multimodal pattern may exist (children, adults, elderly clusters).

FARE BY PASSANGER CLASS



1st Class:

• Has the highest median fare.

• Wide range with many outliers (some extremely expensive tickets).

• Indicates that this class had a mix of upper- and ultra-wealthy passengers.

2nd Class:

• Moderate fare range, tighter distribution than 1st class.

• Still some small outliers, suggesting a few expensive tickets.

3rd Class:

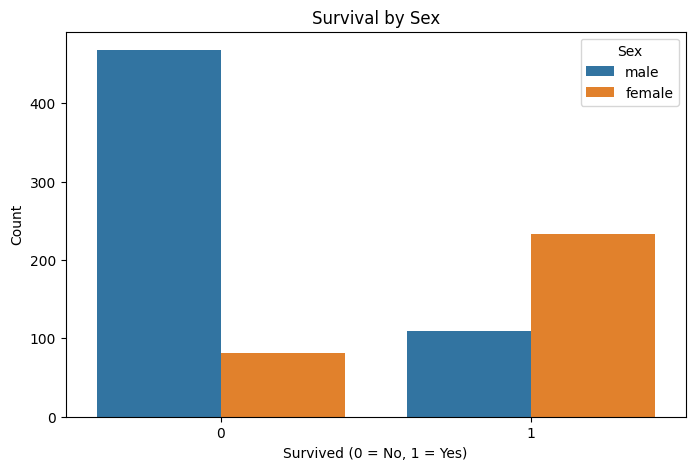
• Lowest fares with very tight distribution.

• Some low outliers indicate very cheap fares, accessible for poorer travelers.

Fare is strongly correlated with class — as expected. The outliers in 1st class show a few passengers

paid significantly more, possibly for private cabins. Visual confirms that class was a proxy for wealth, which likely affected survival chances.

Survival by Sex

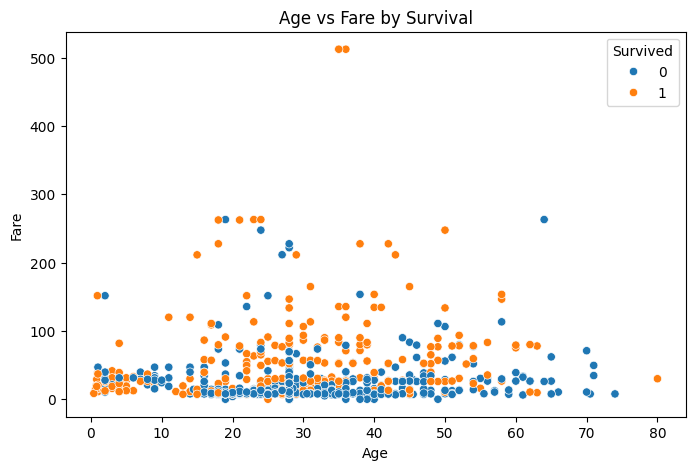
- Most males did not survive.

- Most females survived.

- Female survival rate is significantly higher than male.

- Sex had a strong impact on survival chances.

AGE VS FARE COLORED BY SURVIVAL



- Survivors (Survived = 1, typically shown in red or blue) cluster at:

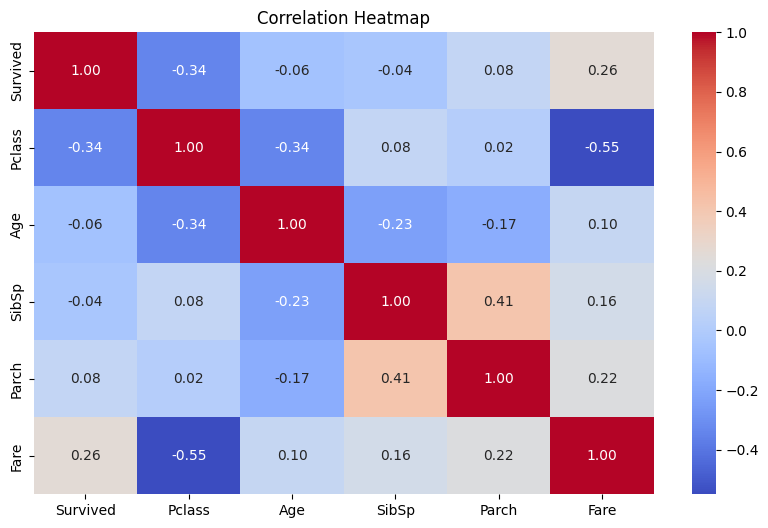
o Higher fares (upper section).

o All age groups, but more young adults and children.

- Non-survivors dominate lower-fare, middle-age groups.

- Some very expensive tickets were paid by both survivors and non-survivors — possibly 1st - class passengers located far from lifeboats or delayed.

CORRELATION HEATMAP



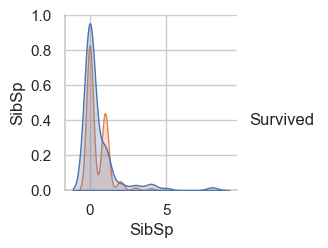
- Fare is positively correlated with Survival (0.26).

- Pclass is negatively correlated with Fare (-0.55) and Survival (-0.34).

- SibSp and Parch show a moderate positive correlation (0.41).

- Other features like Age and Survival have weak or no strong correlation.

DISTRIBUTION OF SIBLINGS/SPOUSES ABOARD (SIBSP) BY SURVIVAL



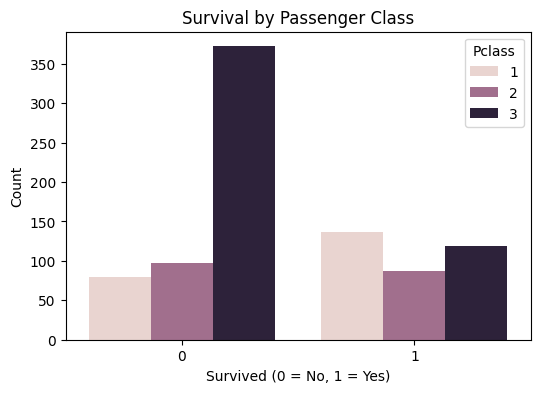
- Most passengers had no siblings/spouses aboard.

- Passengers with 0 or 1 SibSp had higher survival densities.

- Higher SibSp values were less common and linked to lower survival.

- The peak survival density was around SibSp = 0.

SURVIVAL BY PASSENGER CLASS



- The number of people died are more in Class 3. Most probably due to preference, social

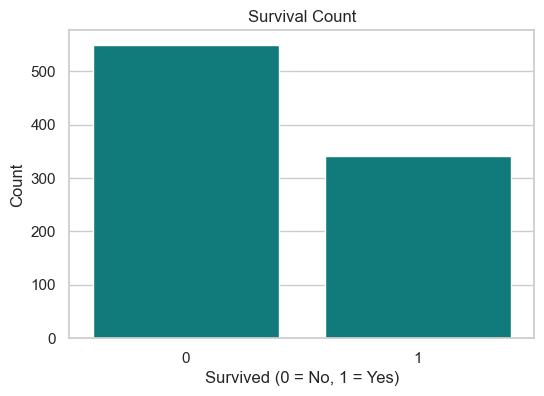
importance and discrimination in that era.

- Similarly, number of people who lived are more in numbers in Class 1. Also, overall count of

people who survived are average in all classes.

- Class played a big role — more access to lifeboats, location of cabins, etc.

SURVIVAL COUNT



- People who survived are nearly 320

- People who died are more in numbers, nearly 550

- Roughly 2/3rd died, 1/3rd survived.

CONCLUSION

There was a significantly higher survival rate among females compared to males, highlighting the impact of the "women and children first" evacuation policy. Passenger class also played a crucial role in survival, with those in first class having much higher chances of survival than those in second or third class. The fare and age distributions further suggest that wealthier and likely older passengers tended to be in higher classes, which correlated with better survival odds. The boxplot and violin plots show that survivors generally paid higher fares, indicating that they belonged to upper classes. Embarkation port analysis showed that most passengers boarded at Southampton, but the survival rate varied across ports. The scatter plot of Age vs Fare showed clustering of survivors in certain fare ranges, especially among younger individuals. The correlation heatmap revealed modest correlations between survival and features like class, fare, and number of siblings/spouses.

Overall, the data and visualizations collectively suggest that gender, socio-economic status, and class were the strongest indicators of survival on the Titanic.