DATA ANALYST INTERNSHIP TASK 5

- Dhairyasen Deshmukh

- Task 5: Exploratory Data Analysis (EDA)
- Objective: Extract insights using visual and statistical exploration.

ABOUT THE DATA:

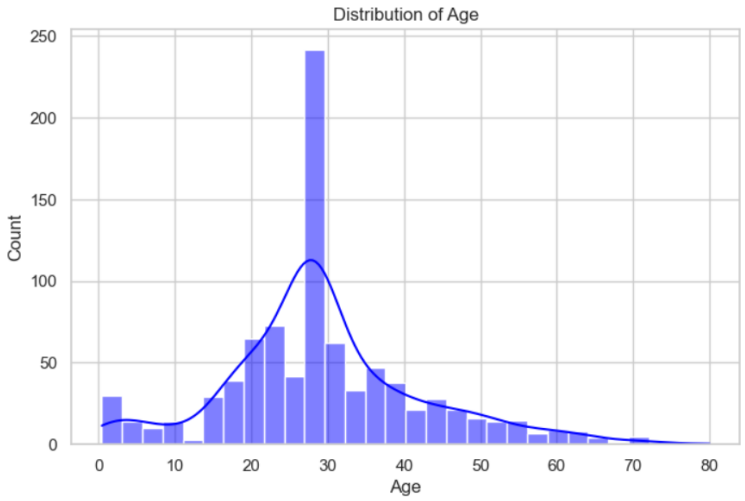
Titanic dataset contains information about passengers aboard the RMS Titanic, with details that can help us analyse patterns—especially regarding survival.

APPROACH TO EXPLORATORY DATA ANALYSIS:

1. Download the dataset from Kaggle.
2. Import the data into Jupyter Notebook.
3. Create a new file named Task5 into the Notebook.
4. Import **pandas, matplotlib, seaborn, openpyxl** into the library.
5. Perform Basic data cleaning on the dataset.
6. 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?
7. Plot charts using imported libraries.
8. Understand Insights of the visualization output.
9. Create a Report of the EDA.

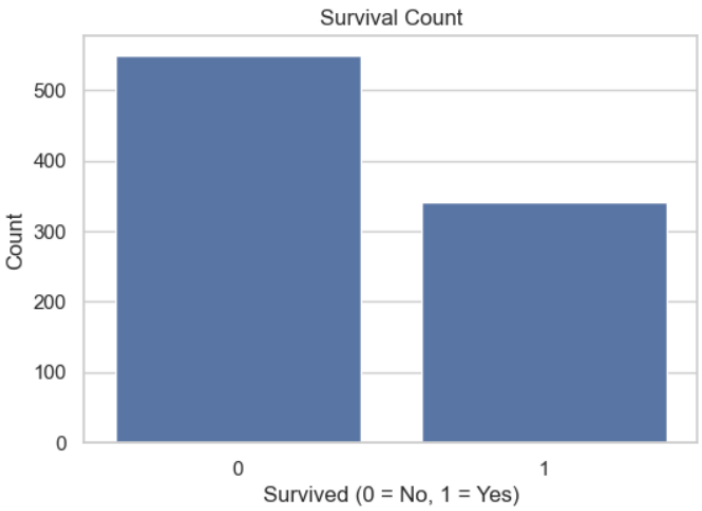
INSIGHTS

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).

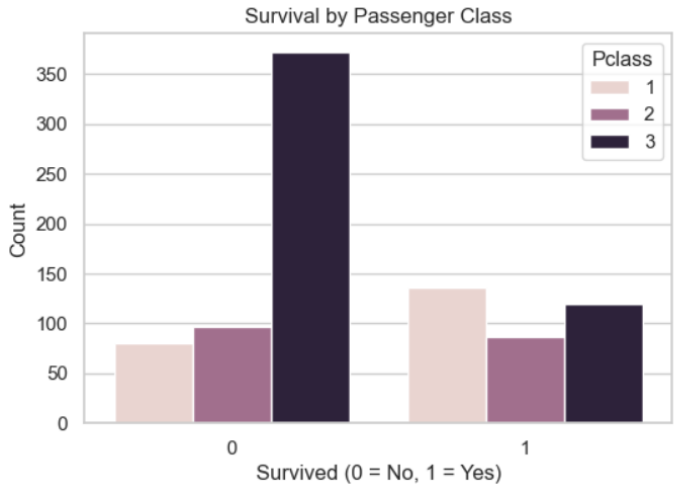
SURVIVAL COUNT



- People who survived are nearly 320
- People who died are more in numbers, nearly 550

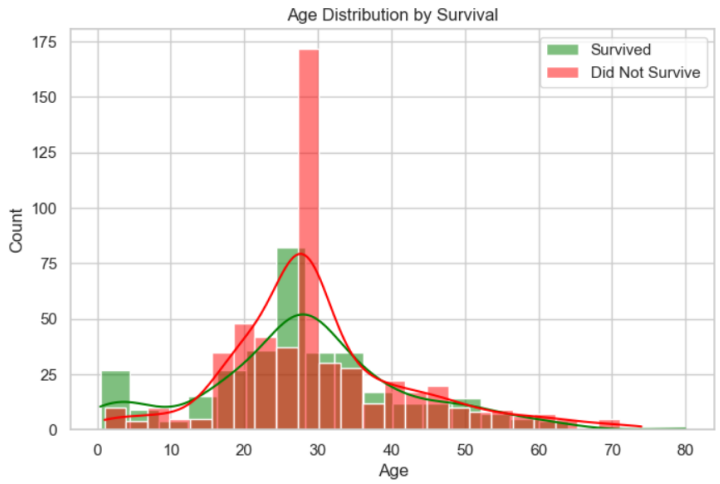
- Roughly 2/3rd died, 1/3rd survived.

SURVIVAL BY PASSANGER CLASS



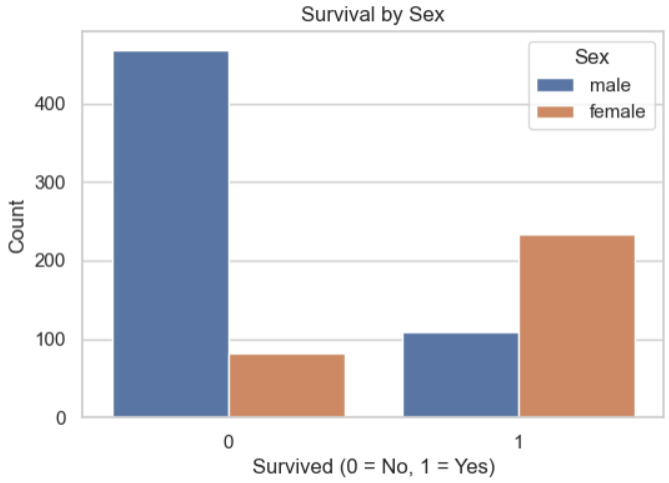
- The number of people died are more in Class 3. Most probably due to preference, social importance and discrimination in that era.
- Similarly, number of people who lived are more in numbers in Class 1. Also, overall count of people who survived are average in all classes.
- Class played a big role — more access to lifeboats, location of cabins, etc.

AGE DISTRIBUTION BY SURVIVAL



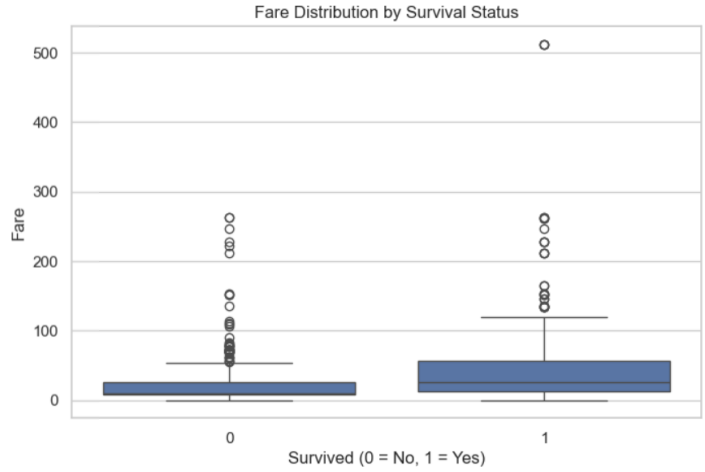
- Number of people who died the most are around the age of 30
- Most likely due to their number being high or volunteering in saving young and old people.

SURVIVAL BY SEX



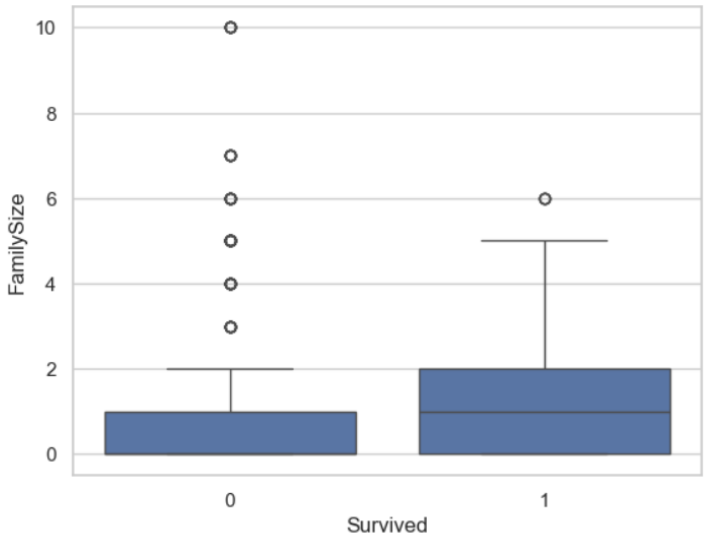
- Females had a much higher survival rate than males.
- Suggests a "women and children first" rescue policy.

FARE DISTRIBUTION BY SURVIVAL



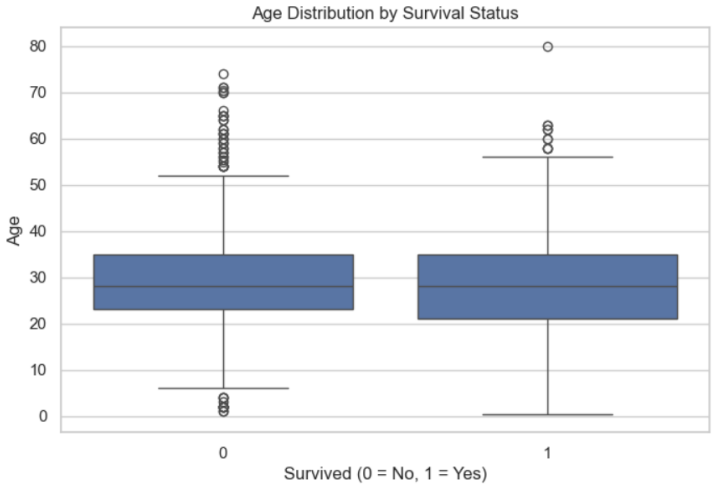
- Survivors paid higher average fares.
- Some outliers show extremely high fares among survivors.
- Suggests wealthier passengers had better chances.

SURVIVAL BY FAMILY SIZE



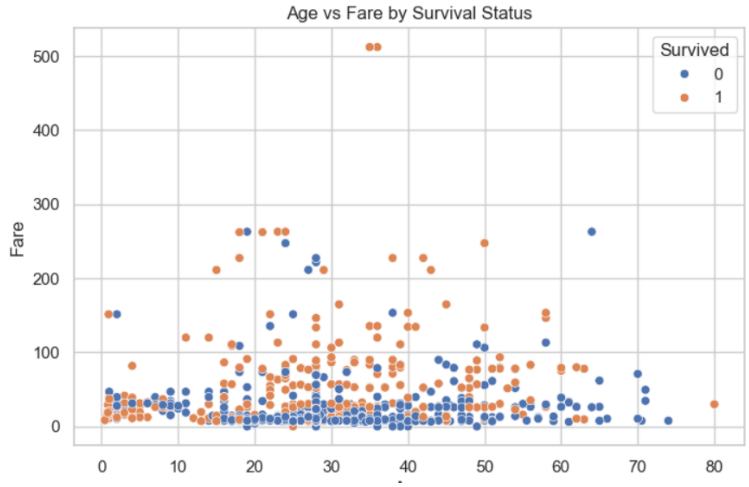
- Passengers with 1-3 family members had higher survival.
- Too large or too small families (0 or >4) had lower survival.

SURVIVAL BY AGE



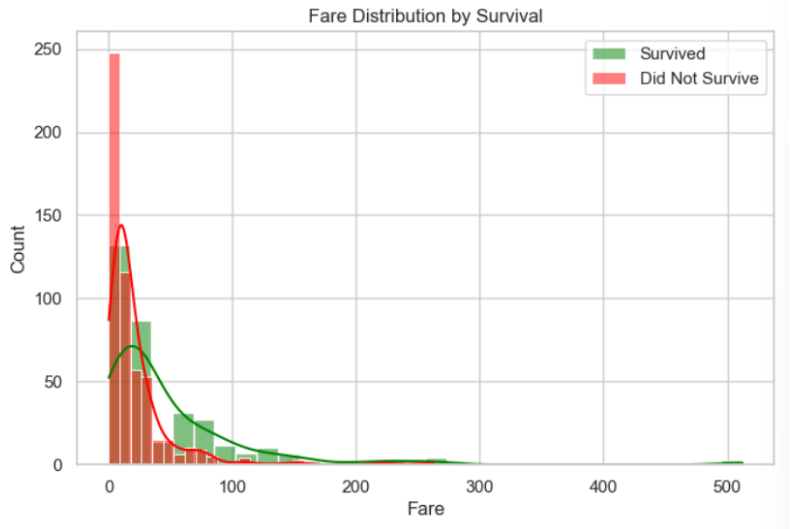
- Survivors were slightly younger on average.
- Some children clearly survived more.
- Elderly had lower chances.

AGE VS FARE BY SURVIVAL RATE



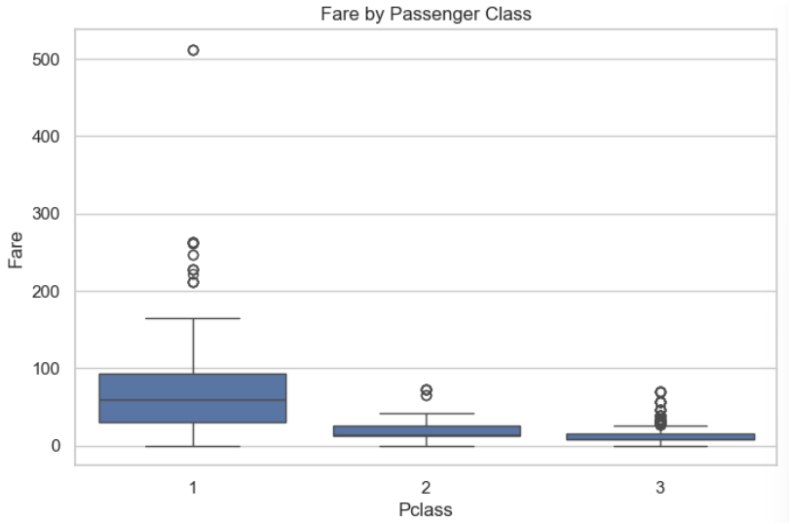
- Survivors (Survived = 1, typically shown in red or blue) cluster at:
    - Higher fares (upper section).
    - All age groups, but more young adults and children.
- Non-survivors dominate lower-fare, middle-age groups.
- Some very expensive tickets were paid by both survivors and non-survivors — possibly 1st-class passengers located far from lifeboats or delayed.

FARE DISTRIBUTION BY SURVIVAL



- Survivors paid higher average fares.
- The trend line clearly shows an upward shift from non-survivors to survivors.
- This supports the idea that wealthier passengers (1st class) had better survival chances.

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 travellers.

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.

CONCLUSION

The Titanic dataset reveals strong patterns in passenger survival based on gender, class, age, and fare. Females had a significantly higher survival rate than males, highlighting the "women and children first" evacuation protocol. Passengers in 1st class, who paid the highest fares, were more likely to survive, while 3rd class passengers with the lowest fares faced the lowest survival rates, showing a clear link between socio-economic status and survival. Most passengers were young adults, but children—especially in higher classes—had better survival chances. Scatter plots showed that younger, higher-paying passengers had better odds of survival, although no strong linear correlation between fare and age was found. Box plots confirmed the stark difference in fare across passenger classes, with 1st class showing the widest fare range and highest median. Overall, survival

was influenced by a combination of gender, class, fare, and age, highlighting inequality in life-and-death outcomes aboard the Titanic.