Exploratory Data Analysis on the Titanic Dataset

1. Introduction:-

The Titanic disaster of 1912 remains one of the most well-documented maritime tragedies in history. More than 1,500 passengers and crew lost their lives after the ship struck an iceberg during its maiden voyage. The dataset derived from passenger records has become a widely used benchmark for practicing data analysis and machine learning, primarily because it contains demographic and travel-related attributes alongside the survival outcome.

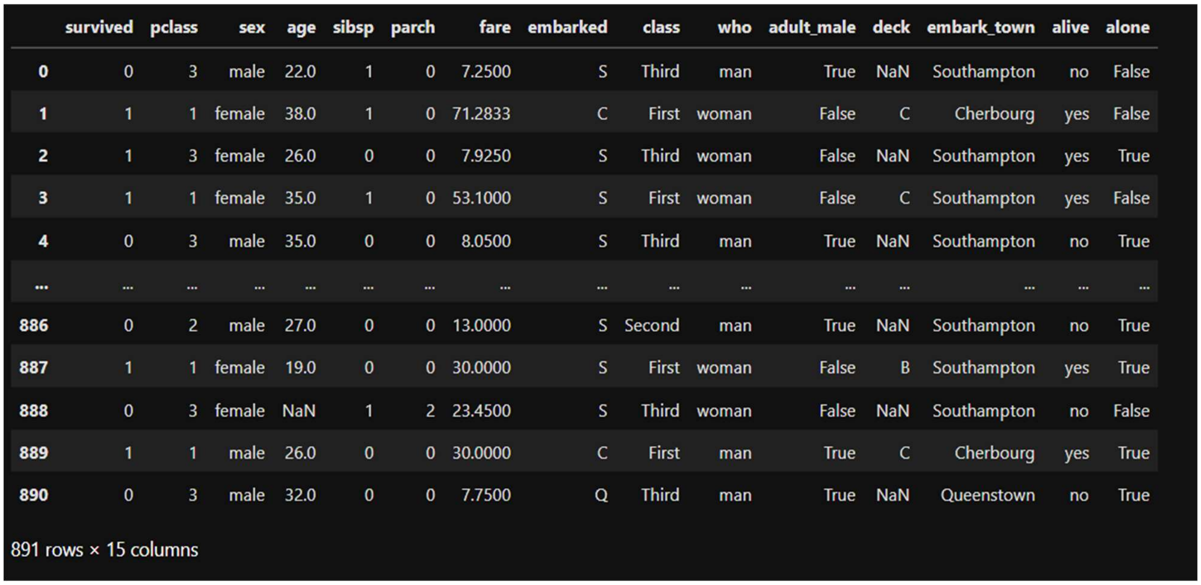
This report presents an exploratory data analysis (EDA) of the Titanic dataset. EDA is a critical first step in any data science workflow, allowing analysts to uncover patterns, detect anomalies, test hypotheses, and form intuitions that guide predictive modelling. By combining descriptive statistics and visualizations, we aim to understand the underlying structure of the data and identify key factors influencing survival.

2. Objective:-

The main objectives of this analysis are:

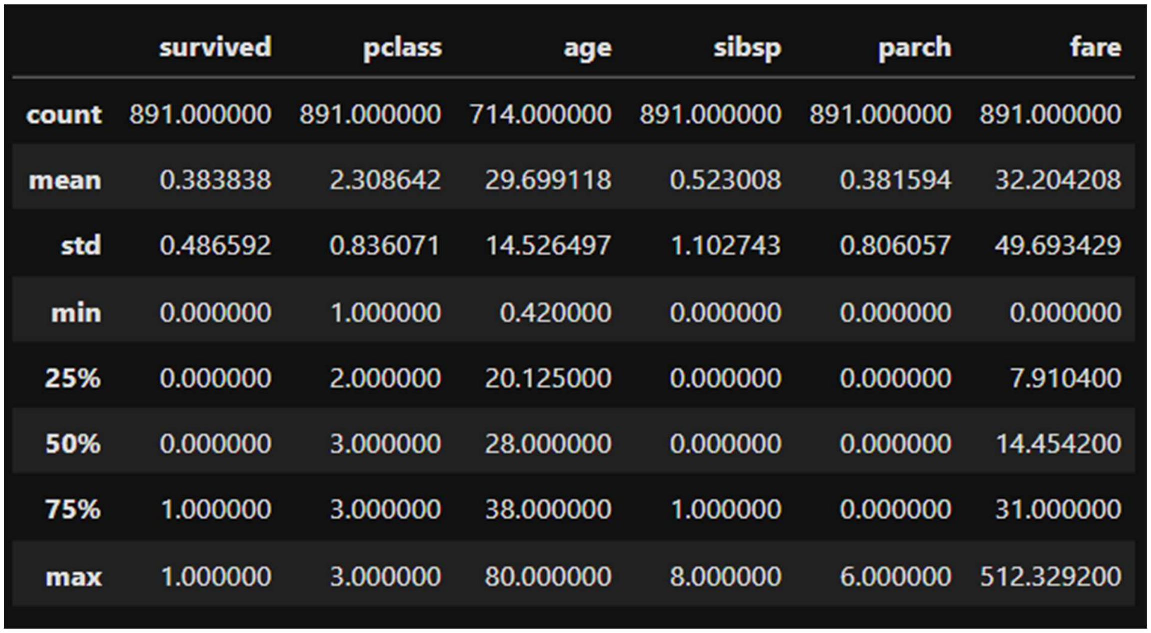
- To explore the dataset and summarize its key characteristics.
- To handle missing values, outliers, and skewed variables appropriately.
- To examine relationships between passenger demographics, travel attributes, and survival.
- To visualize important patterns and trends through histograms, boxplots, scatterplots, heatmaps, and pairplots.
- To document findings and generate insights that can inform feature engineering and predictive modelling.

3. Overview of the dataset:-



	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	Second	man	True	NaN	Southampton	no	True
887	1	1	female	19.0	0	0	30.0000	S	First	woman	False	B	Southampton	yes	True
888	0	3	female	NaN	1	2	23.4500	S	Third	woman	False	NaN	Southampton	no	False
889	1	1	male	26.0	0	0	30.0000	C	First	man	True	C	Cherbourg	yes	True
890	0	3	male	32.0	0	0	7.7500	Q	Third	man	True	NaN	Queenstown	no	True

891 rows × 15 columns



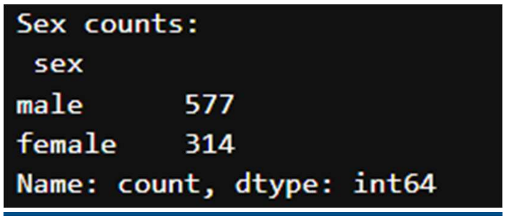
	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Observations:

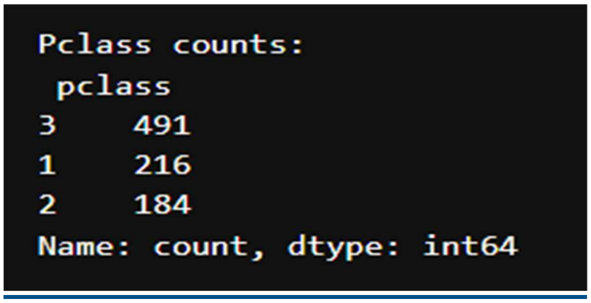
- Dataset has 891 rows and 15 columns.

- Variables like Age and Cabin have missing values.

- 'sex' and 'pclass' are imbalanced (more males, most passengers in class 3).

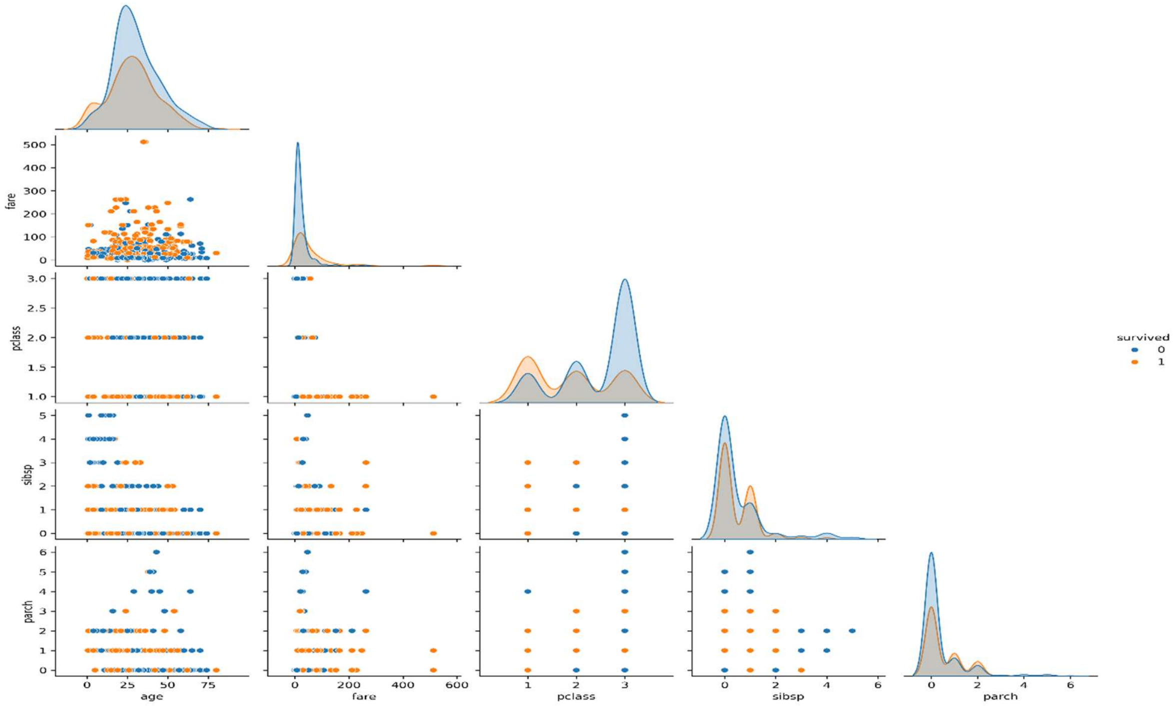


```
Sex counts:
 sex
male      577
female    314
Name: count, dtype: int64
```

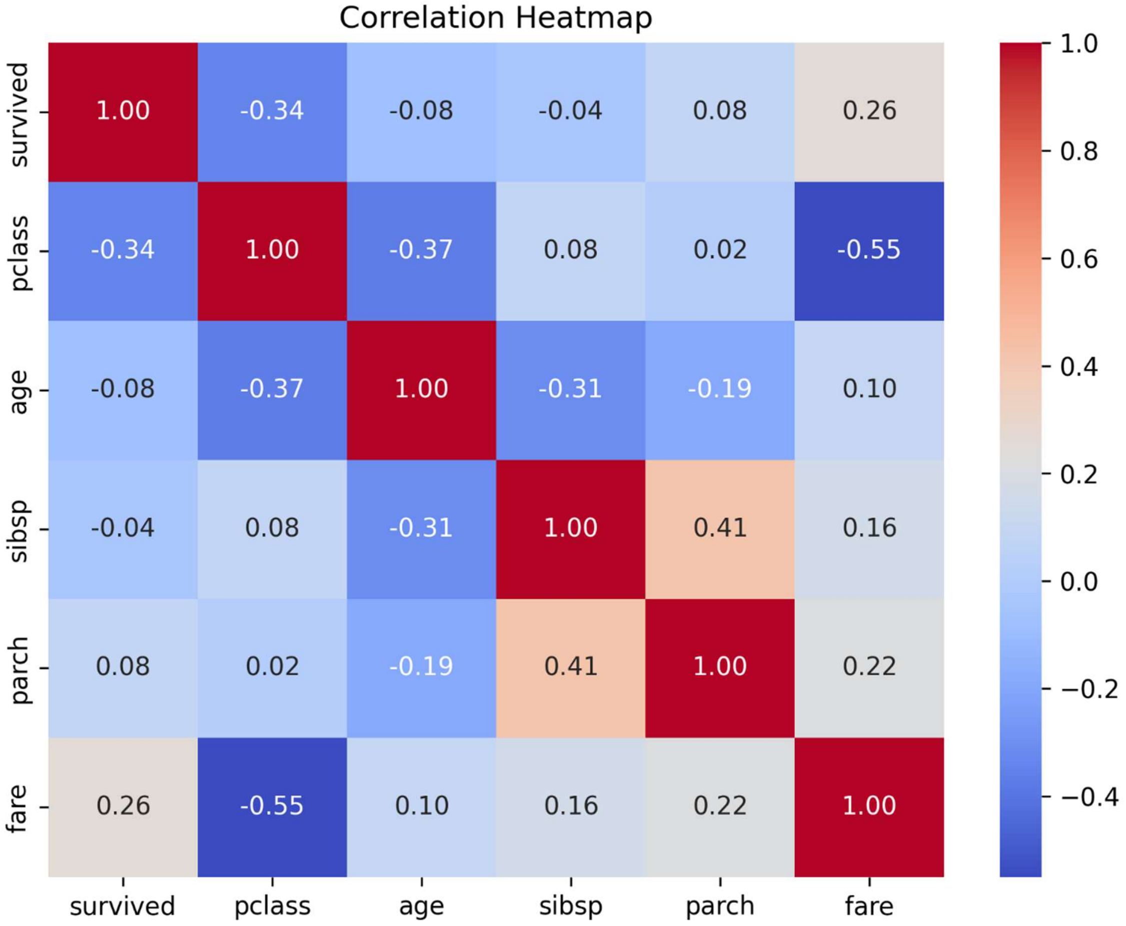


```
Pclass counts:
 pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

Pairplot shows clear class separation in 'Fare' and 'Pclass' by survival.

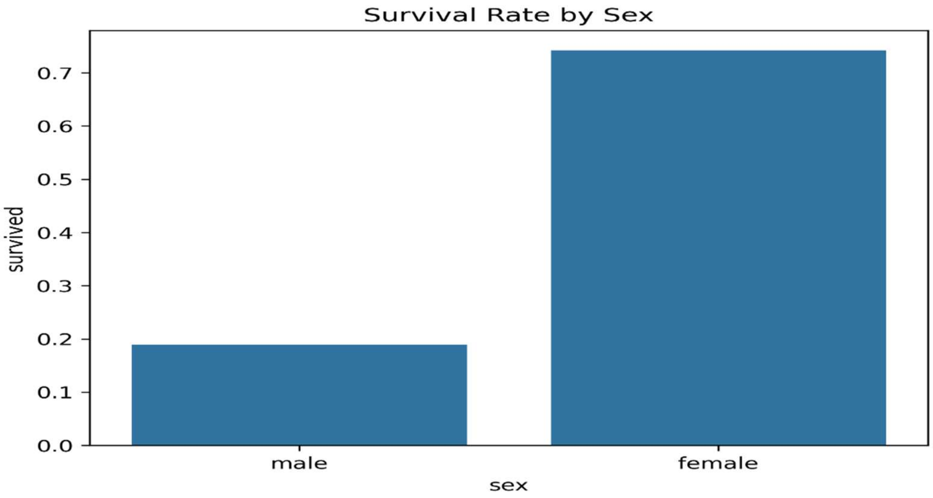


Heatmap: Fare is negatively correlated with Pclass; Family features (SibSp, Parch) show weak correlation.

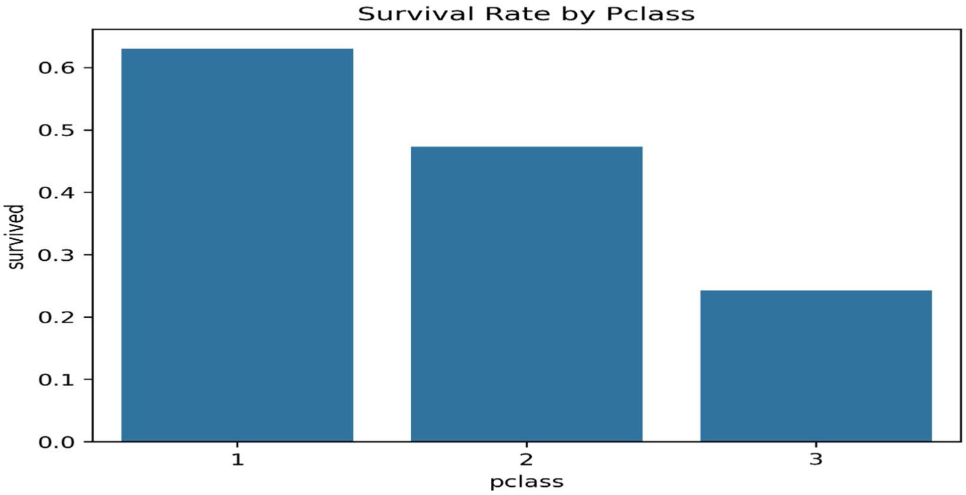


Relationships and Trends

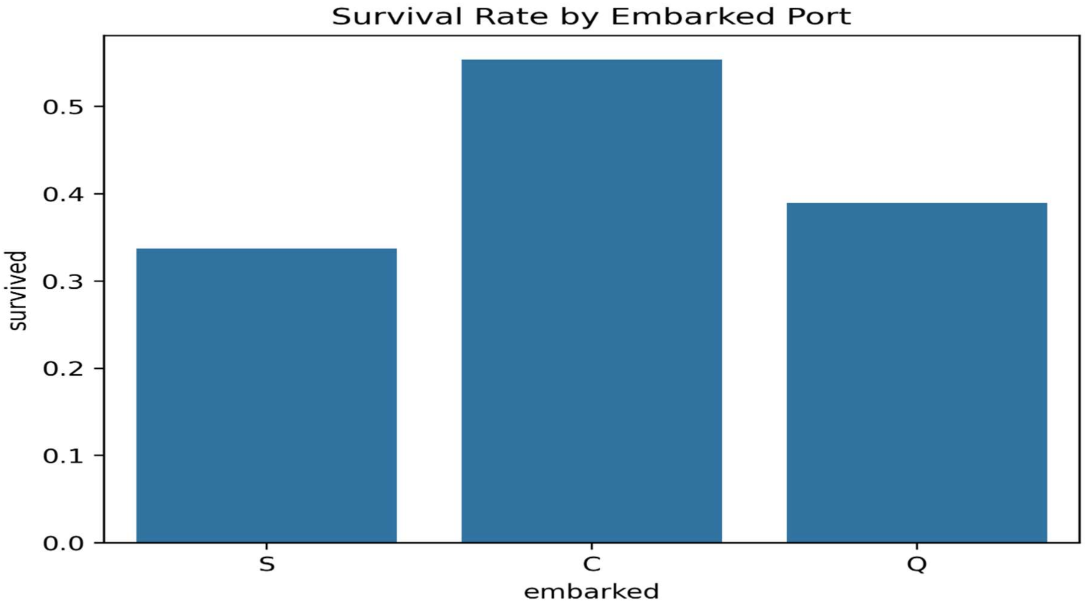
- Females survived at much higher rates than males.



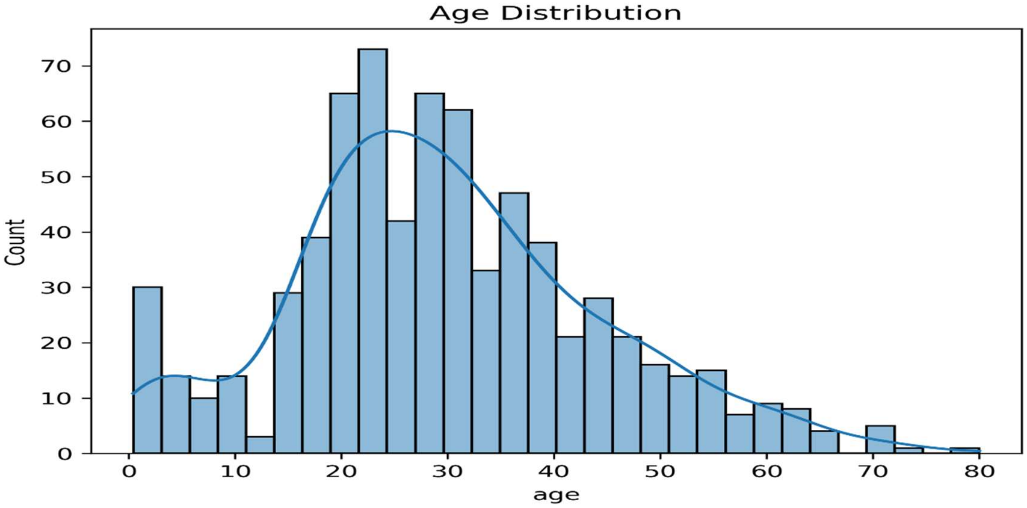
- 1st class passengers had much higher survival than 3rd class.



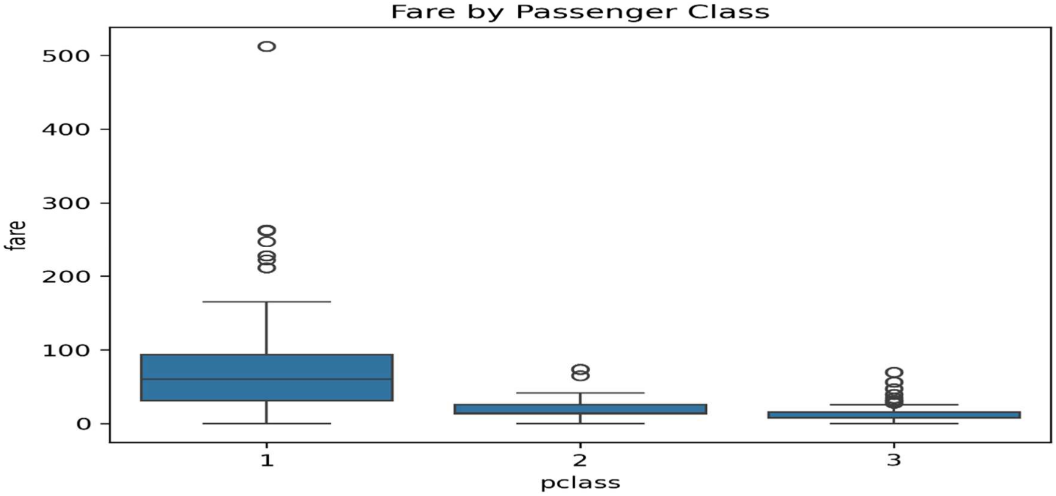
- Passengers from port 'C' had higher survival than 'S'.



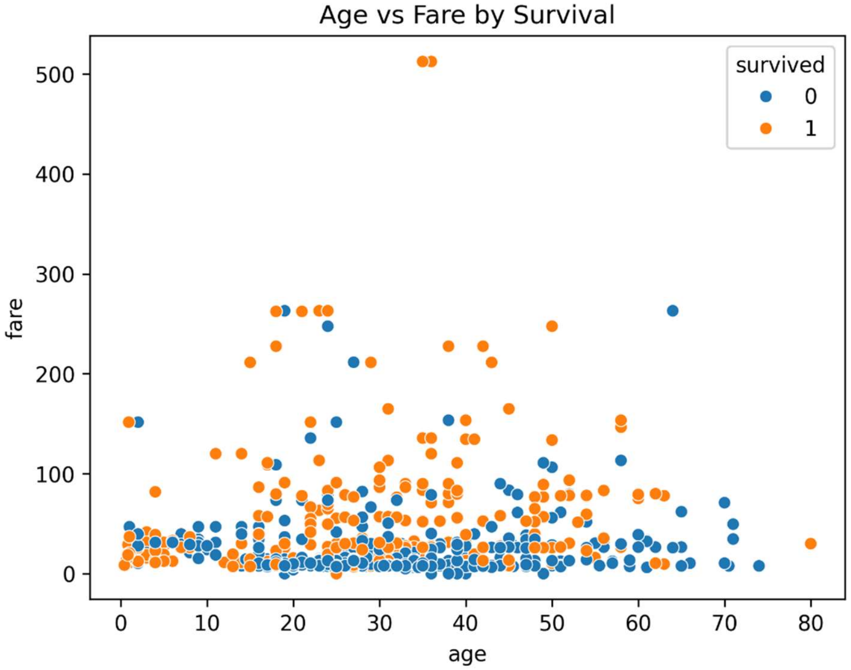
- Age distribution is right-skewed with many young adults.



- Fare is much higher in 1st class with large variance (outliers).



- Survivors cluster at higher fares; some young children in 3rd class also survived.



Summary of Findings

- **Demographics** Majority were male, traveling in 3rd class, age distribution centered around 20–30.
- **Survival Patterns**
- Females and children had significantly higher survival rates.
- 1st class passengers had better outcomes compared to 3rd class.
- Passengers embarking at Cherbourg (C) showed higher survival.
- **Relationships**
- Fare and class strongly linked; higher fares → higher survival.
- Family size had mixed effect: small families better than large or alone.
- **Data Issues**
- Missing values in Age and Cabin handled by imputation/feature engineering.
- Fare was skewed, log-transform reduces skewness.
- **Next Steps**
- Use engineered features (Family Size, Title, Deck) in predictive models.
- Drop highly collinear features if needed (e.g., SibSp & Parch vs Family Size).

Conclusion:-

The exploratory data analysis of the Titanic dataset revealed several important insights into the factors influencing passenger survival. Demographic attributes such as sex and age, alongside travel-related features such as class and fare, showed clear associations with survival outcomes. Females, children, and first-class passengers exhibited higher survival rates, while third-class male passengers had the lowest survival probabilities.

The analysis also highlighted data quality issues, including missing values in Age and Cabin, as well as the presence of skewness in the Fare variable. Appropriate handling through imputation, feature engineering (e.g., family size, passenger title, deck extraction), and transformation improved the dataset's suitability for further modelling.

Overall, the EDA not only provided descriptive insights into passenger survival patterns but also laid the foundation for predictive modelling by identifying key predictors and potential multicollinearity concerns.

Recommendations:-

□ Feature Engineering: Retain engineered features such as *Family Size*, *Title*, and *Deck*, as they capture survival-related information beyond the raw variables.

□ Model Readiness: Use transformed versions of skewed variables (e.g., log-transformed Fare) and consider dropping or combining collinear variables.

□ Future Analysis: Extend the study with predictive modelling (e.g., logistic regression, decision trees, or ensemble models) to quantify survival probabilities.

□ Visualization: Include additional plots (e.g., survival by family size group, survival by combined sex and class) to further strengthen insights.