🚢Who Lived and Why : Exploratory Data Analysis on Titanic Dataset

🧭 1. Introduction

The Titanic dataset offers a unique glimpse into the passengers aboard the ill-fated ship. This project explores survival patterns using univariate, bivariate, and multivariate analysis while applying various feature engineering and data transformation techniques to enrich insights and modeling potential.

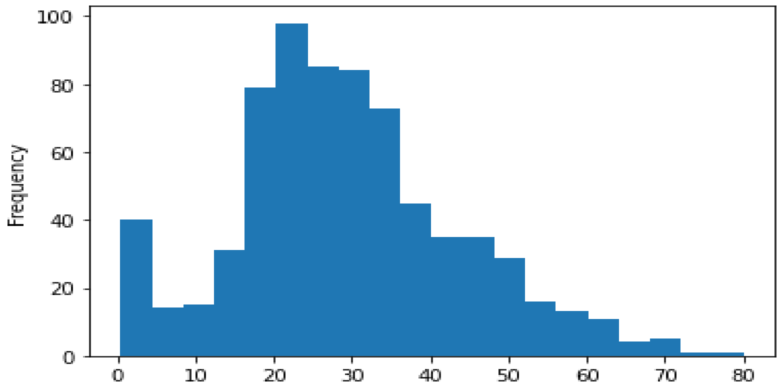
📂 2. Data Overview

The dataset includes features such as:

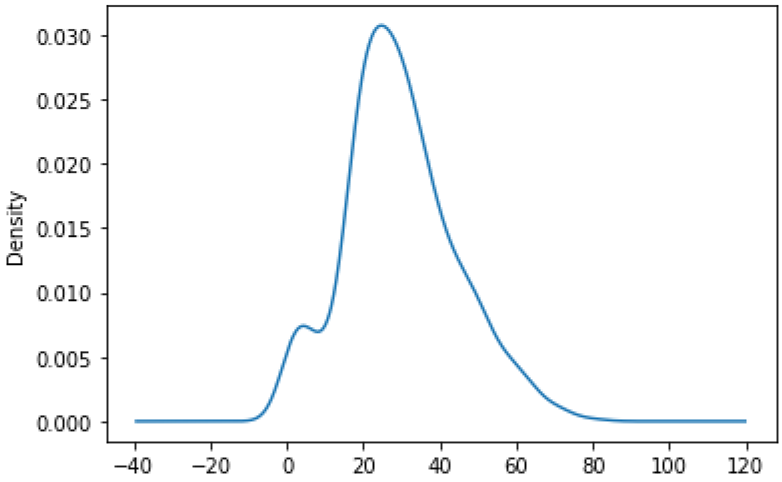
- **Demographics**: Age, Sex, Class
- **Family structure**: SibSp, Parch
- **Ticket info**: Fare, Cabin, Embarked
- **Target variable**: `Survived` (1 = yes, 0 = no)

📈 3. Univariate Analysis & Insights

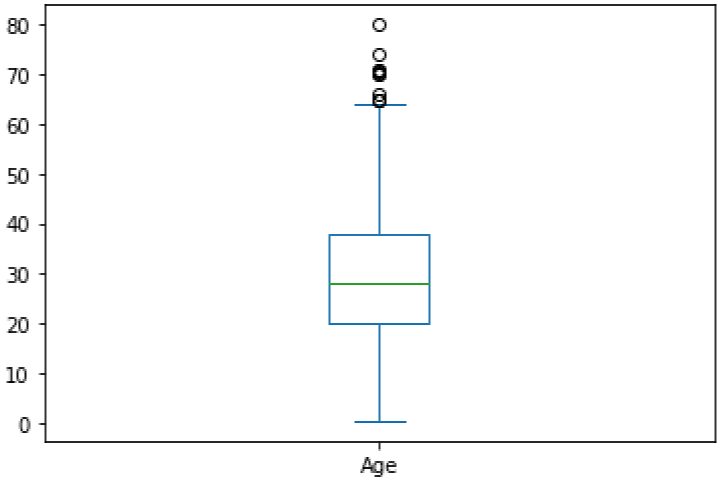
🔹 Age Distribution



- **Histogram**: Shows the age concentration between 20–40 years.



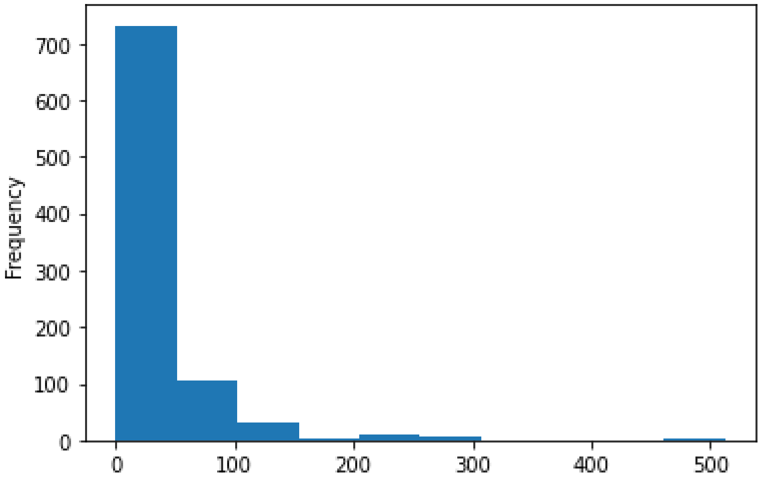
- **KDE Plot**: Indicates a near-normal distribution, with a slight right skew (`skewness ≈ 0.39`).



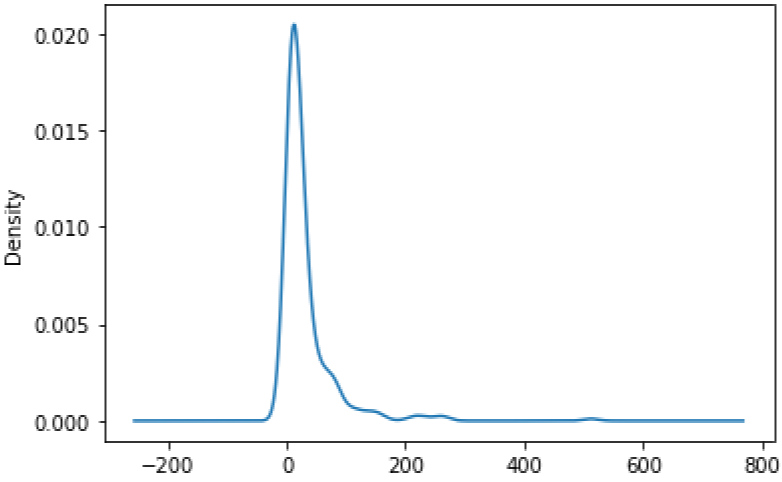
- **Boxplot**: Highlights outliers above 65; most passengers are younger.
- **Missing Values**: 19.87% of `Age` values are missing.
-

📌 **Insight**: The majority were young adults; few children or elderly. This may impact survival probability.

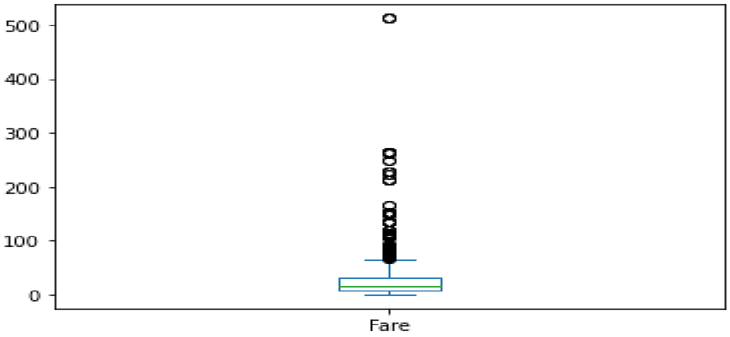
🔹 Fare Distribution



- **Histogram**: Right-skewed distribution with many low-fare passengers.



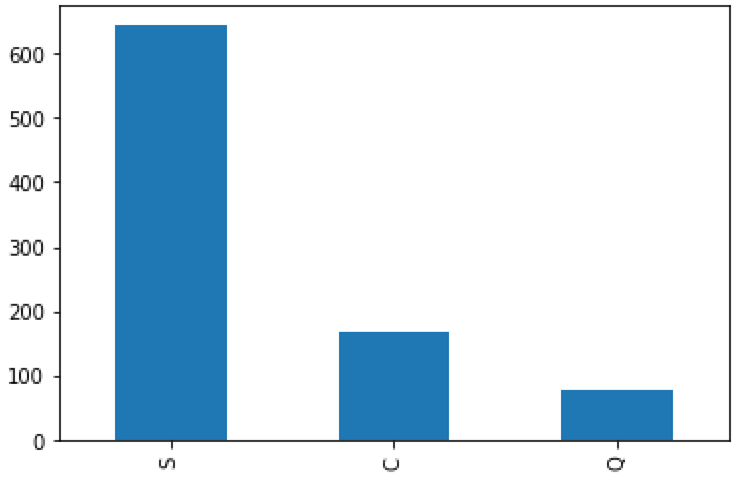
- **KDE Plot**: Strong positive skew (`skewness ≈ 4.79`), indicates long tail.



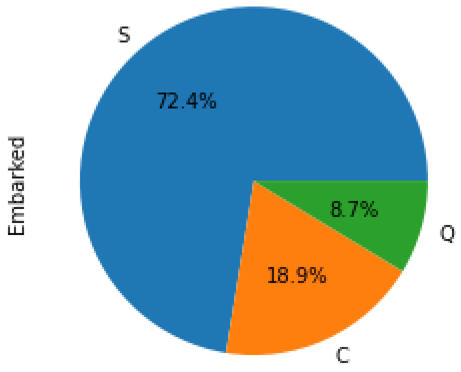
- **Boxplot**: Outliers above 250, especially high-class passengers.

📌 **Insight**: Most passengers paid <50 fare units; extreme values from wealthy families inflated the mean.

🔹 Embarked



- **Bar Chart**: Most boarded at Southampton (S), followed by Cherbourg (C), and Queenstown (Q).



- **Pie Chart**: S = 72.4%, C = 18.9%, Q = 8.7%

📌 **Insight**: Embarkation port could indirectly affect survival—S passengers were more numerous in lower class.

🔄 4. Feature Engineering

🧠 `individual_fare`

```
df['individual_fare'] = df['Fare'] / (df['SibSp'] + df['Parch'] + 1)
```

✔️ Adjusted group fare to per-person.
📌 **Outcome**: Reduced skew; median individual fare more aligned with personal-level analysis.

👨‍👩‍👧‍👦 `family_size` and `family_type`

```
df['family_size'] = df['SibSp'] + df['Parch'] + 1

def transform_family_size(num):
    if num == 1: return 'alone'
    elif num < 5: return 'small'
    else: return 'large'

df['family_type'] = df['family_size'].apply(transform_family_size)
```

✔️ Defined group travel behavior.
📌 **Outcome**: Small families had best survival rates; large families faced worse outcomes (~84% non-survival).

🏷️ `title` and `surname`

```
df['surname'] = df['Name'].str.split(',').str.get(0)
df['title'] = df['Name'].str.split(',').str[1].str.strip().str.split(' ').str.get(0)
```

✔️ Parsed titles and grouped rare ones as `'other'`.

📌 **Outcome**: Survival rates varied by title. E.g., `Mrs.` and `Miss.` had higher chances; `Mr.` fared worse.

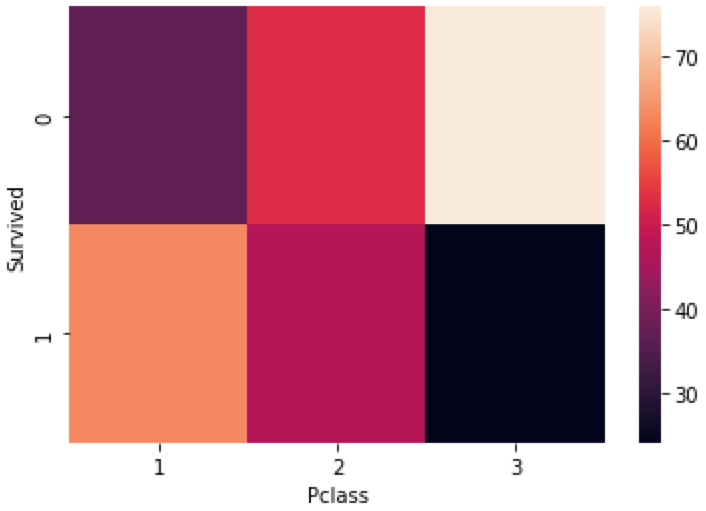
🧭 deck

```
df['Cabin'].fillna('M', inplace=True)
df['deck'] = df['Cabin'].str[0]
```

✔️ Extracted the first letter to represent deck level.
📌 **Outcome**: Deck A/B/C correlated with higher survival; Deck M (unknown cabins) dominated lower classes.

🔗 5. Bivariate Analysis & Insights

📌 Pclass vs Survival



- **Heatmap**: 1st class → 62% survival, 3rd class → 24% 📌 **Insight**: Strong class disparity; privilege likely influenced rescue priority.

📌 Sex vs Survival

```
pd.crosstab(df['Survived'], df['Sex'], normalize='columns') * 100
```

- Females: 74% survived
- Males: 19% survived

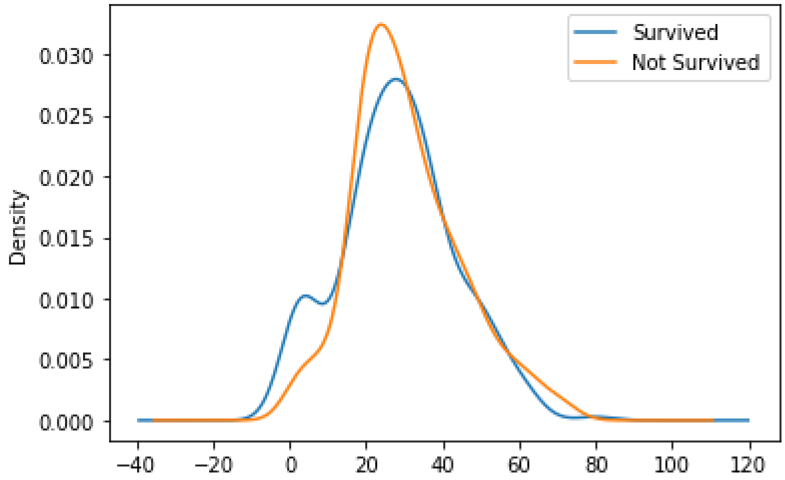
> 📌 **Insight**: "Women and children first" was followed during evacuation.

📌 Embarked vs Survival

- Cherbourg (C): 55% survived
- Queenstown (Q): 39%
- Southampton (S): 34%

  📌 **Insight**: Embarked location indirectly hints at class and survival odds.

📌 Age vs Survival (KDE)



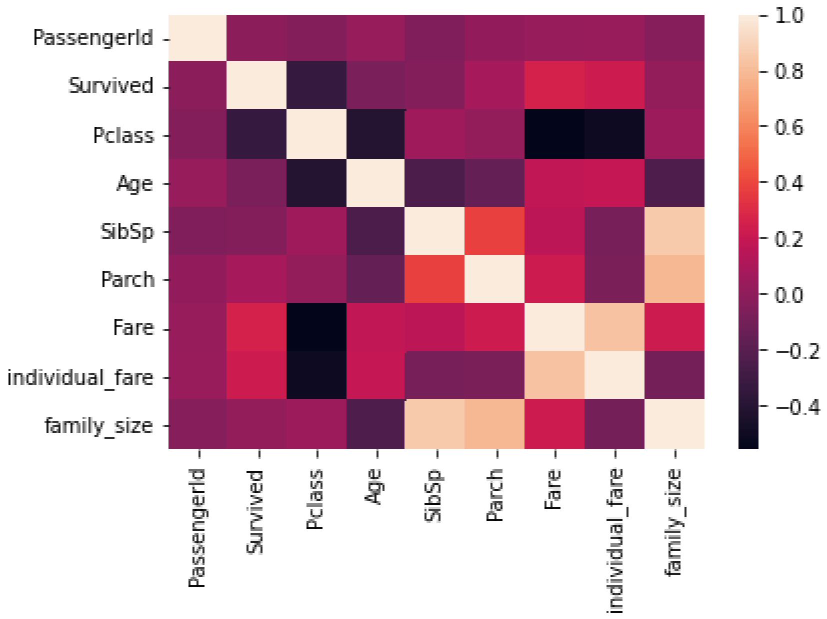
Overlayed plots show children and young adults had marginally better survival rates.

⚙️ 6. Challenges & Solutions

Problem	Solution
Missing `Age`, `Cabin`	Imputed `Cabin` with 'M'; `Age` flagged and engineered around titles
Complex `Ticket`/`Cabin` fields	Ignored sparse tickets, simplified `Cabin` into `deck`
Skewed fare	Derived `individual_fare`; reduced extreme influence
Title diversity	Grouped similar rare titles into `'other'`

🔍 7. Correlation Heatmap

- **Survived** positively correlated with `Fare`, `individual_fare`, and inversely with `Pclass`.



- Weak or no correlation with `PassengerId`, `SibSp`.

✅ 8. Conclusion : Survival Dependency on Key Attribute

Understanding how each variable influenced survival provided critical insights for hypothesis generation and future modeling. Here's a breakdown of the most impactful attributes:

🔸 Sex

- **Observation**: ~74% of females survived, compared to only ~19% of males.
- **Explanation**: This strong disparity suggests the evacuation protocol prioritized women, aligning with maritime traditions like “women and children first.”

🔸 Pclass (Passenger Class)

- **1st Class**: ~62% survival
- **2nd Class**: ~47% survival

- **3rd Class**: ~24% survival
- **Insight**: Higher-class passengers likely had better cabin placement and quicker lifeboat access, directly improving survival odds.

🔸 Age

- Younger passengers (especially children) had modestly better survival rates, as seen in KDE overlays.
- Older passengers (>65) were fewer and showed lower survival likelihood.
- **Dependency**: Age is a weak-to-moderate predictor when isolated but gains power when cross-referenced with other features (e.g., class, sex).

🔸 Fare and `individual_fare`

- Higher fare correlates with higher survival.
- Feature `individual_fare` helped remove family-size bias and revealed more consistent patterns.
- **Explanation**: High-fare tickets were associated with better decks and boarding priority.

🔸 Embarked Port

- **Cherbourg (C)**: Highest survival (≈55%)
- **Queenstown (Q)** and **Southampton (S)**: Lower survival
- **Interpretation**: These differences mirror class proportions—Cherbourg had more 1st-class passengers.

🔸 Family Type

- **Small families** (2–4 people): ~58% survival
- **Alone**: ~30%
- **Large families** (>5): ~16%
- **Insight**: Moderate group size may have helped coordination during evacuation; large groups struggled.

Deck

- Decks A, B, and C (upper decks): Higher survival
- Deck M (missing cabin data): Lowest rates, linked to 3rd class
- **Conclusion**: Cabin placement played a key spatial role.

Title

- Titles like **Miss.** and **Mrs.** had higher survival; **Mr.** had the lowest.
- Grouping rare titles (e.g., Dr., Col., Rev.) into an 'other' category revealed nuanced social hierarchies.