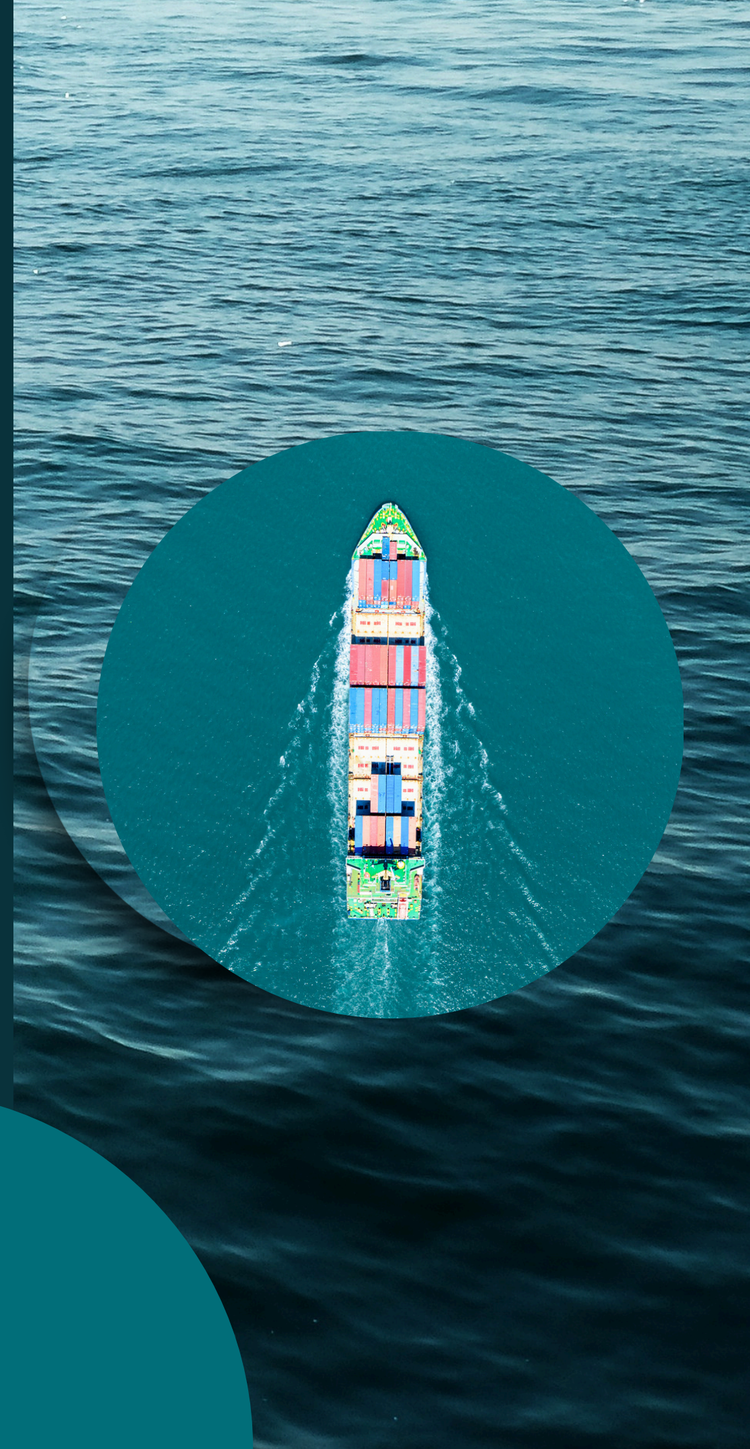
EXPLORATORY DATA ANALYSIS ON TITANIC DATASET

by Marchelia Zafira

START PRESENTATION



Introduction

The RMS Titanic sank on April 15, 1912, after striking an iceberg, resulting in the deaths of 1502 out of 2224 passengers and crew. This tragedy shocked the world and exposed major flaws in maritime safety regulations, particularly the insufficient number of lifeboats onboard.

Beyond the technical shortcomings, the disaster also raised critical questions about who survived and why. What factors influenced survival rates? Was it age, gender, social status, or combination of these ?

Project Goals

- Clean and preprocess the Titanic dataset to ensure data quality and consistency
- Conduct Exploratory Data Analysis (EDA) to explore distribution and correlation
- Extract strategic insights from the visual analysis to better understand passenger characteristics and survival factors

Dataset Overview

survived	name	sex	age
1	Allen, Miss. Elisabeth Walton	female	29
1	Allison, Master. Hudson Trevor	male	1
0	Allison, Miss. Helen Loraine	female	2
0	Allison, Mr. Hudson Joshua Creighton	male	30
0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25

```
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   survived  500 non-null    int64
 1   name      500 non-null    object
 2   sex       500 non-null    object
 3   age       451 non-null    Int64
```

The dataset comprises 500 rows and 4 columns containing information about Titanic passengers.

Data Preprocessing

Duplicate Handling

```
len(data.drop_duplicates()) / len(data)
# Ensure the output of this cell is 1. If it's not, duplicate records exist in the dataset
```

```
0.998
```

The output is not 0, which means there are duplicate values

```
# Step 1: Extract all duplicate rows, including the original entries
duplicates = data[data.duplicated(keep=False)]
duplicates
```

	survived	name	sex	age
104	1	Eustis, Miss. Elizabeth Mussey	female	54
349	1	Eustis, Miss. Elizabeth Mussey	female	54

The output shows 2 identical rows

```
#Remove duplicate entries and verify that no duplicates remain
data = data.drop_duplicates()
len(data.drop_duplicates()) / len(data)
```

```
1.0
```

Remove duplicate by using drop_duplicates function. The output returns 1, which means the duplicate value has been handled

Missing Value Handling

```
#Count the number of missing values in each column
data.isna().sum()
```

	0
survived	0
name	0
sex	0
age	49

- There is no missing value in column survived, name, and sex
- Column age has 49 missing value

```
# percentage version
total_rows = len(data)

# Count percentage of missing values in each column
for column in data.columns:
    missing_count = data[column].isna().sum()
    missing_percentage = (missing_count / total_rows) * 100
    print(f"Column '{column}' Has {missing_count} missing values ({missing_percentage:.2f}%)")
```

```
Column 'survived' Has 0 missing values (0.00%)
Column 'name' Has 0 missing values (0.00%)
Column 'sex' Has 0 missing values (0.00%)
Column 'age' Has 49 missing values (9.82%)
```

- The percentage of missing values is below 20%, so we keep the columns and handle numeric columns using the median and categorical columns using the mode. However, the categorical columns do not contain any missing values.

Missing Value Handling

```
data['age'].median()
data['age'].fillna(data['age'].median())
```

- Calculates median of age column and fills the missing values with median

```
data.isna().sum()
```

	0
survived	0
name	0
sex	0
age	0

- The output confirms that there are no remaining missing values

```python
def preprocess_data(data2):
    # Extract titles from the 'name' column
    data2['title'] = data['name'].str.extract(r',\s*([^\.]*)\s*\.')

    # Creating Binary Columns for Each Title
    titles = data2['title'].unique()
    for title in titles:
        data2[f'title_{title}'] = (data2['title'] == title).astype(int)

    # Convert 'sex' to binary (1 for male, 0 for female)
    data2['sex_binary'] = (data2['sex'] == 'male').astype(int)

    # Categorize age into bins
    bins = [0, 4, 9, 18, 59, 100]
    labels = ['Toddler', 'Child', 'Teenager', 'Adult', 'Senior']
    data2['age_category'] = pd.cut(data2['age'], bins=bins, labels=labels)

    return data2

# Run preprocessing
data2 = preprocess_data(data)

# Check the result
data2.info()
data2.sample(5)
```

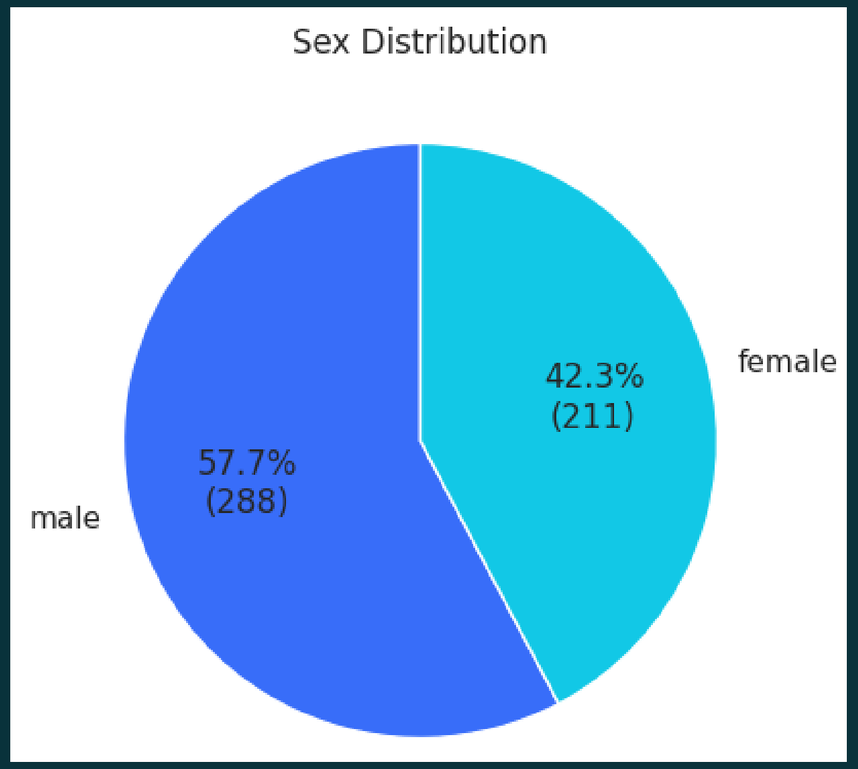
- Extracts titles from name column to identify social status, which may influence survival rate
- The code creates binary columns for each unique title in the title column
- Converts sex to binary, making it suitable for statistical analysis (correlation)
- Groups age to simplify analysis and highlight survival patterns across age ranges

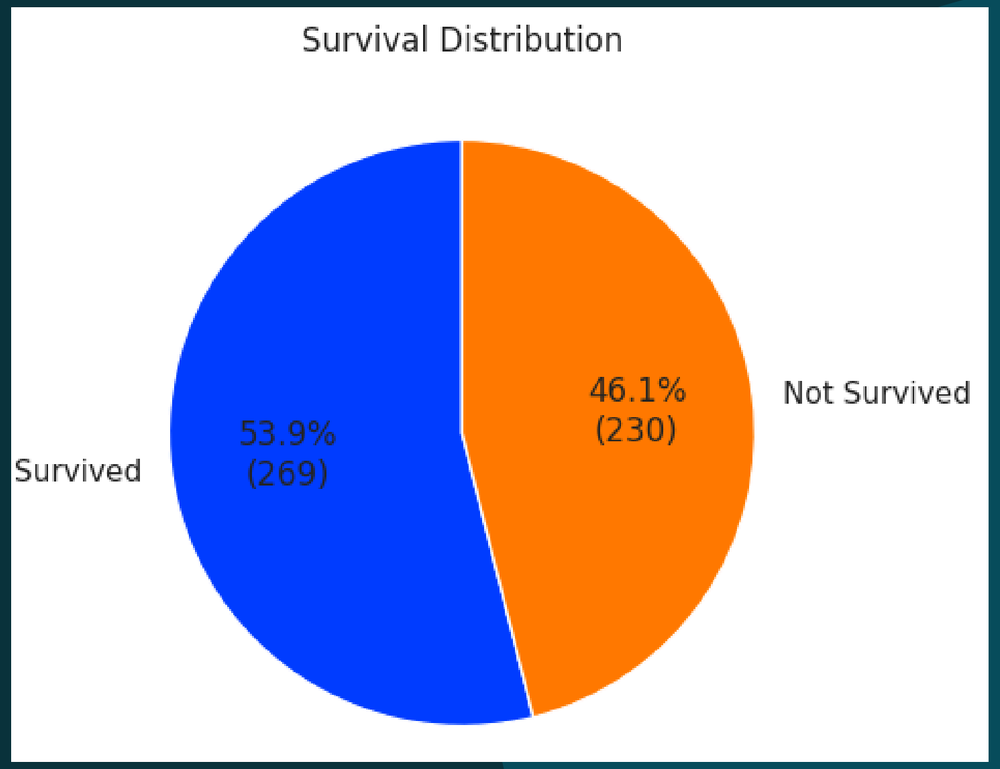
Final Dataset

```
Index: 499 entries, 0 to 499
Data columns (total 24 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   survived             499 non-null    int64
 1   name                 499 non-null    object
 2   sex                  499 non-null    object
 3   age                  499 non-null    Int64
 4   title                499 non-null    object
 5   title_Miss           499 non-null    int64
 6   title_Master         499 non-null    int64
 7   title_Mr             499 non-null    int64
 8   title_Mrs            499 non-null    int64
 9   title_Col            499 non-null    int64
 10  title_Mme            499 non-null    int64
 11  title_Dr             499 non-null    int64
 12  title_Major          499 non-null    int64
 13  title_Capt           499 non-null    int64
 14  title_Lady           499 non-null    int64
 15  title_Sir            499 non-null    int64
 16  title_Mlle           499 non-null    int64
 17  title_Dona           499 non-null    int64
 18  title_Jonkheer       499 non-null    int64
 19  title_the Countess   499 non-null    int64
 20  title_Don            499 non-null    int64
 21  title_Rev            499 non-null    int64
 22  sex_binary           499 non-null    int64
 23  age_category         499 non-null    category
```

Exploratory Data Analysis

Sex & Survival Distribution

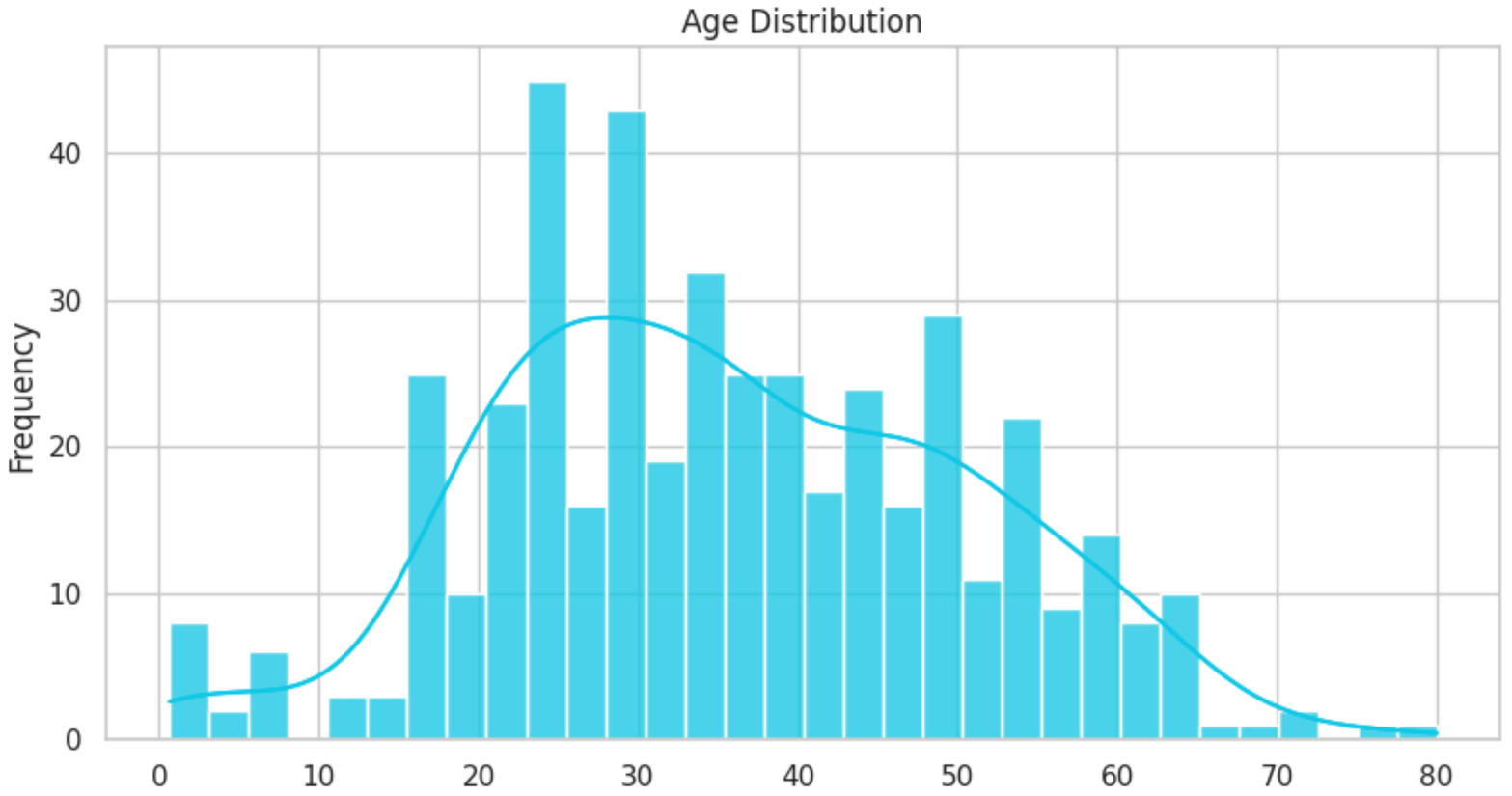




This indicates a gender imbalance, with male passengers representing a larger portion

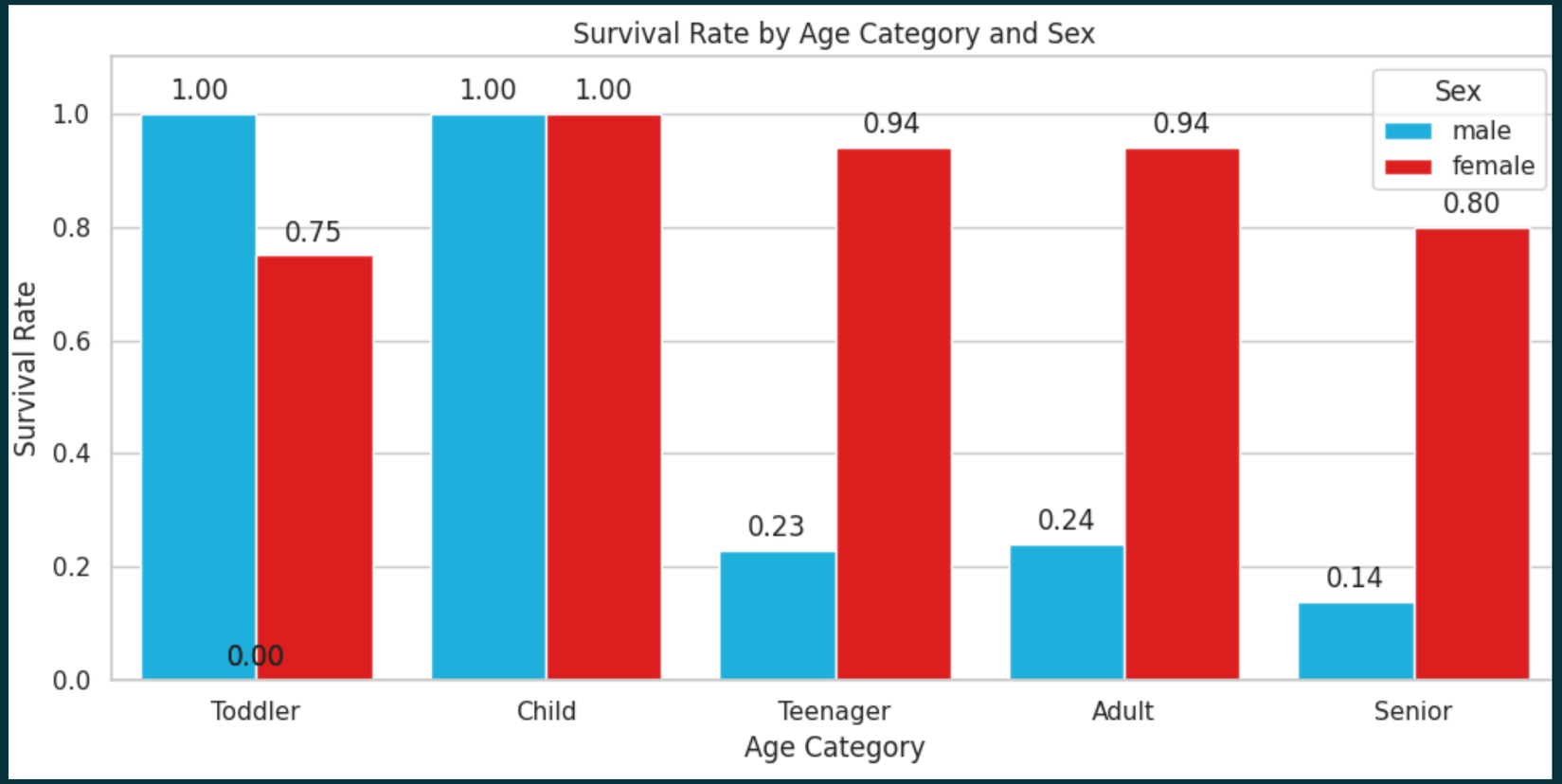
This implies that while survival rates were marginally better, the number of people who did not survive was still relatively high, and the overall outcome remained quite close.

Age Distribution



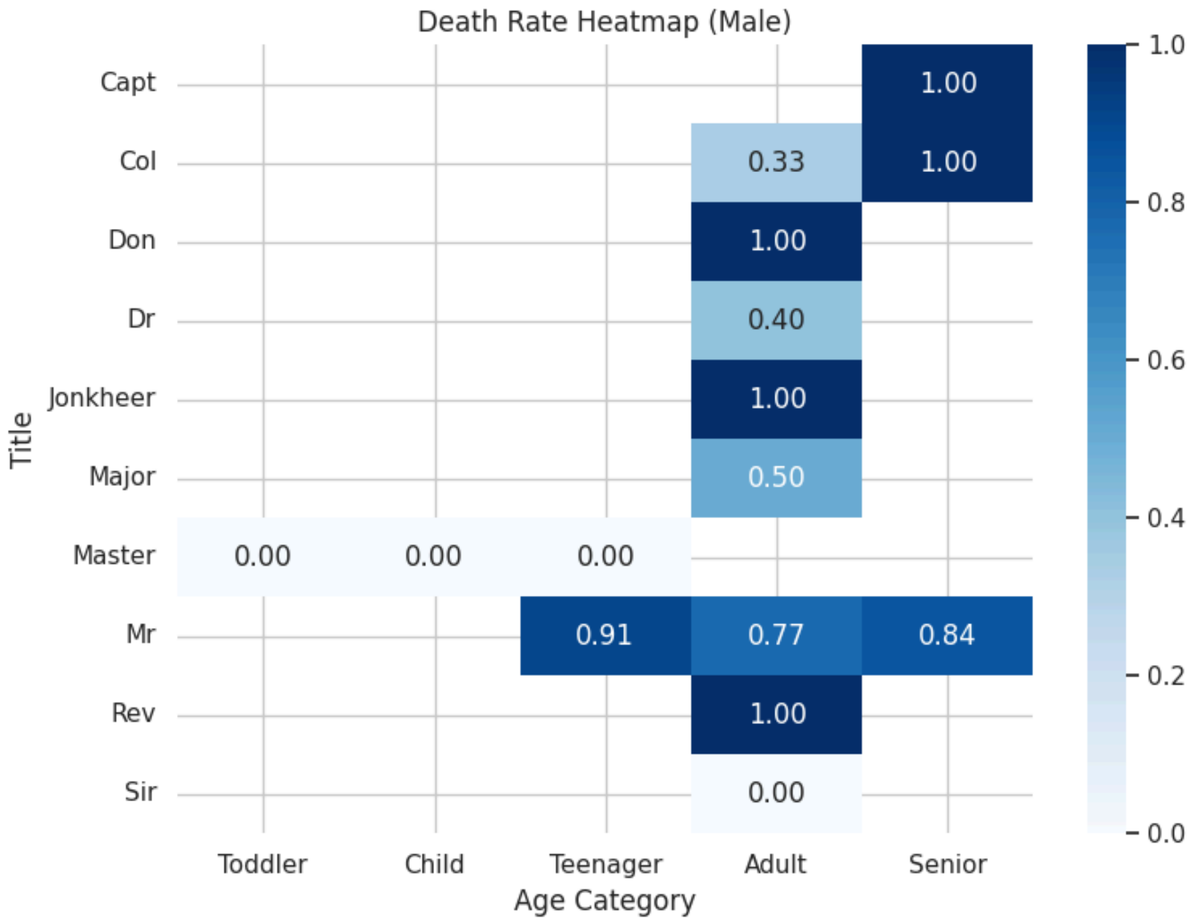
- The majority of passengers were between the ages of 20 and 40, with the highest number around 24 years old.
- There were also a few very young passengers, some even under the age of 5.
- After 40, the frequency of passengers gradually declined, with only a small proportion being over 60 years old.
- The distribution curve is right-skewed, showing that younger passengers were more frequent than older ones.

Survival Rate by Age Category and Sex



- In general, female had higher survival rate.
- When looking at survival rate by age category, toddler and child both male and female had higher survival rate than others.
- These striking disparities highlight the critical role gender and age played in the survival of passengers.
- The pattern reflects the "women and children first" policy that was implemented during the Titanic disaster. The policy prioritized the safety of women and children, which likely contributed to the higher survival rates among these groups

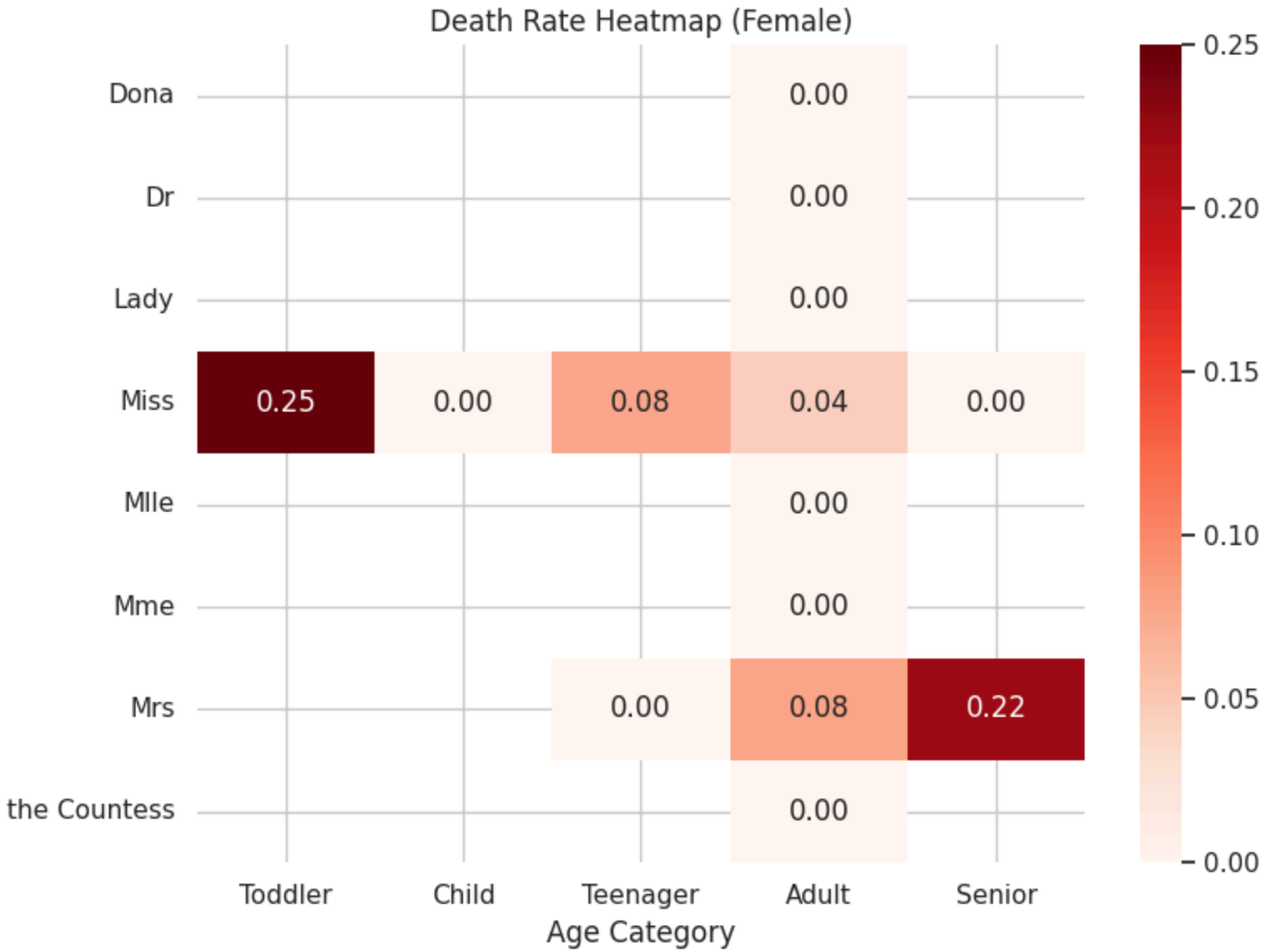
Death Rate Heatmap (Male)



- **High-Risk Titles:**
  - Titles like Capt, Rev, Col (Senior) all show a 100% death rate. This suggests male passengers with official titles had very low survival chances
  - Mr had consistently high death rates indicating that ordinary adult males were the most vulnerable group.
- **Low-Risk Titles:**
  Master (typically used for boys) had 0% death rate across all ages, showing young male children were prioritized for rescue.

- **Interesting Contrasts:**
  - Sir had a 0% death rate, suggests that socially respected may have received special treatment and were likely in first class which had easier access to lifeboats
  - In contrast, Jonkheer and Don had 100% death rate, implying that title alone did not guarantee safety

Death Rate Heatmap (Female)

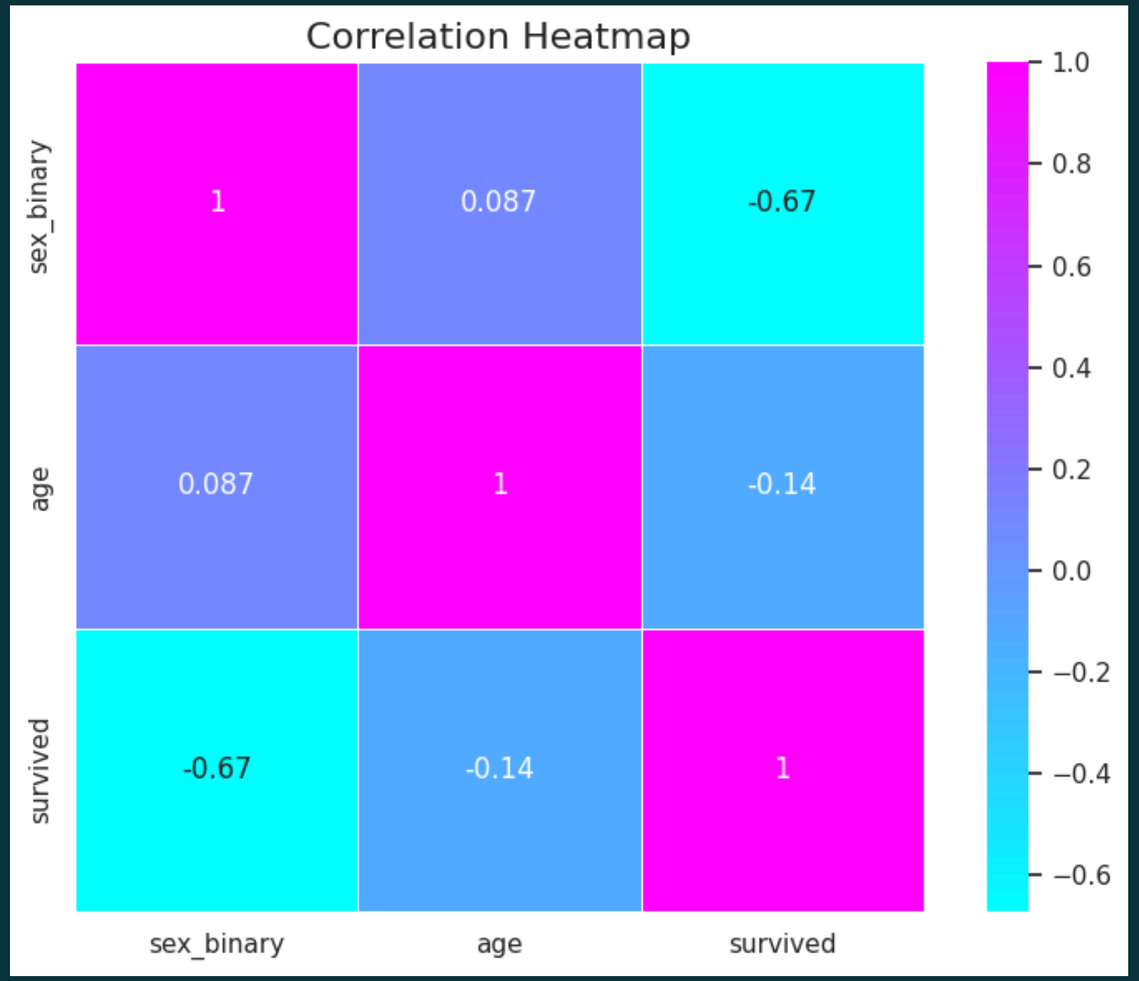


Most female titles show near-zero death rates across age groups, confirming that female passengers were heavily prioritized.

Senior Mrs had the highest death rate among females. Probably because of the physical limitations in emergencies or fewer seats available later during evacuation.

Noble women such as Lady, Dona, and the Countess had complete survival, possibly reflecting both their gender and upper-class status so they may have had quicker access to lifeboats.

Correlation Heatmap



Sex vs Survived = -0.67

- This strong negative correlation shows that males had much lower survival chances.
- Females were given priority for lifeboats, which significantly increased their survival rates.

Sex vs Age = 0.087

- This very weak positive correlation suggests that there was no meaningful relationship between age and sex.
- Both male and female passengers had similar age distributions, indicating that gender differences in survival were not due to age

Survived vs Age = -0.14

- This weak negative correlation suggests younger passengers had a slightly higher chance of survival, supporting the idea that children were prioritized during evacuation.

Conclusion

Gender - key factor: Female passengers had significantly higher survival rates than males, likely due to the "women and children first" evacuation policy.

Age - moderate factor: Toddlers and children—regardless of gender—were more likely to survive compared to adults.

Social status - contributing factor: Passengers with noble or high-status titles (like Lady, Dona, Countess, & Sir) had high survival rate but not all noble titles survived, some like Don and Jonkheer had 100% death rates, showing that status alone wasn’t enough.

THANK YOU