TITANIC SURVIVAL

Uncovering Insights from the Data

BY SYIFA AZZAHRA

DIBIMBING.ID

Introduction

The Titanic was a British passenger liner that tragically sank on its maiden voyage in April 1912 after hitting an iceberg in the North Atlantic Ocean. This disaster led to the loss of more than 1,500 lives, making it one of the deadliest maritime tragedies in history. Much research has been conducted to understand the survival rates of passengers and identify the factors that influenced who survived the disaster.

This project analyzes the Titanic dataset, which includes passenger details like name, survival status, age and gender. Through exploratory data analysis (EDA), I aim to uncover patterns and understand factors affecting survival. Using Python libraries like Pandas, Matplotlib, and Seaborn, I will clean the data, handle missing values, and visualize key insights into Titanic survival.

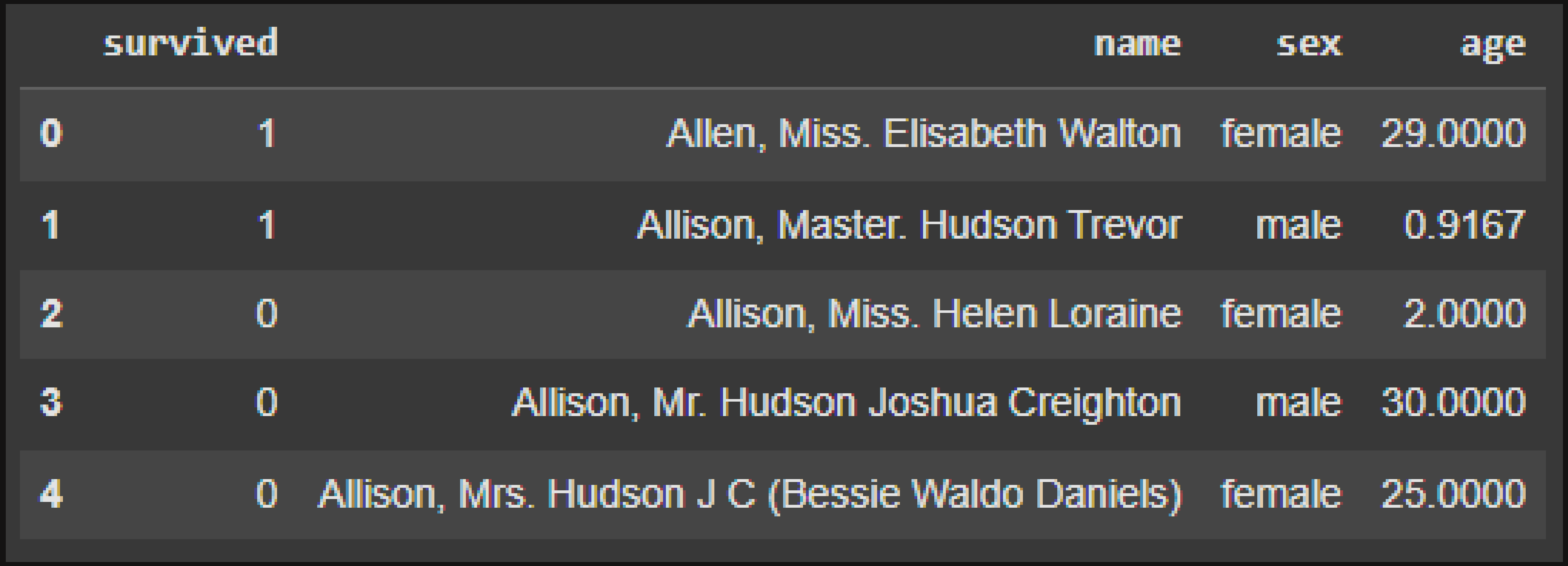
Project Overview

Objective

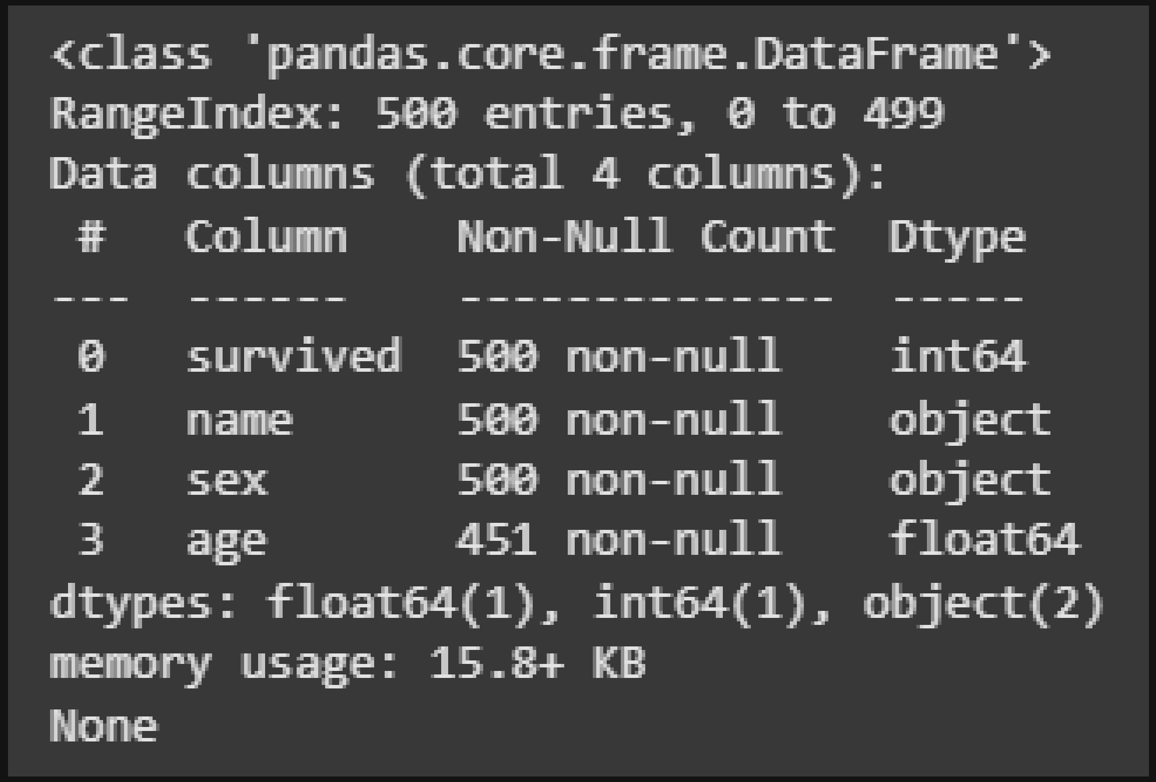
Analyze and explore the Titanic dataset to identify patterns and insights related to passenger survival.

Key Steps

- *Data Observation*
- *Statistical Summary and Initial Data Visualization*
- *Data Preprocessing*
- *Final Data Visualization and Analysis Results*



	survived	name	sex	age
0	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
2	0	Allison, Miss. Helen Loraine	female	2.0000
3	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
4	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000



```
<class 'pandas.core.frame.DataFrame'>
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    float64
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
None
```

Data Observation

Data Observation is the step in data analysis to understand the dataset structure, identify the existing columns, and check for issues such as missing values or duplicates. In this step, a quick look at the data is taken to get an overview of its content and quality.

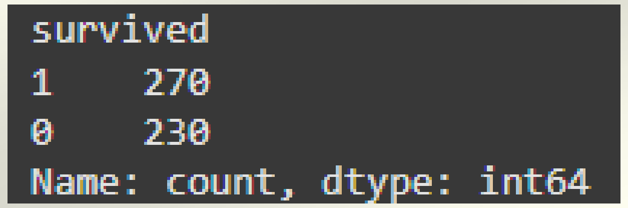
The observation results show that this dataset consists of 500 rows and 4 columns: survived (survival status), name (name), sex (gender), and age (age). The sex column contains two categories (male, female), and survived indicates whether the passenger survived (1) or not (0). The age column has 49 missing values (NaN).

Statistical Summary and Initial Data Visualization

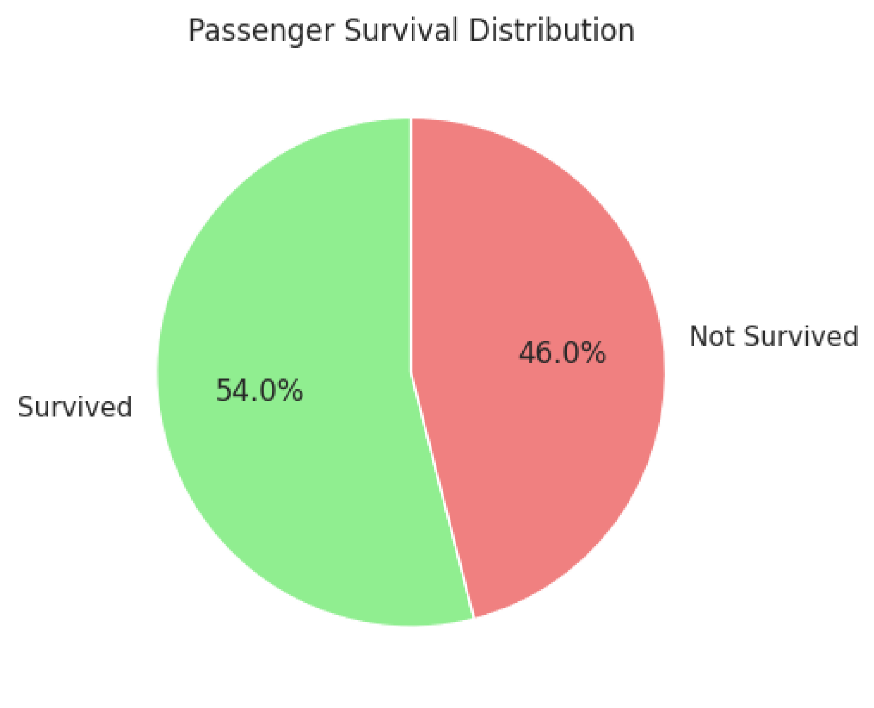
At this step, basic statistics for numerical data, such as mean, standard deviation, and minimum/maximum values, are calculated to understand the data distribution and key characteristics. For categorical data, frequency counts or the mode are used to understand the distribution and the most common categories in the dataset.

Initial visualizations like pie charts and bar plots are used to depict the data distribution. These visualizations help identify key patterns and facilitate further analysis.

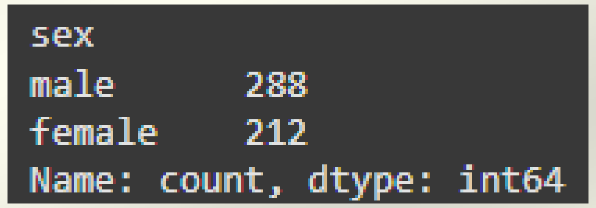
PASSENGER SURVIVAL



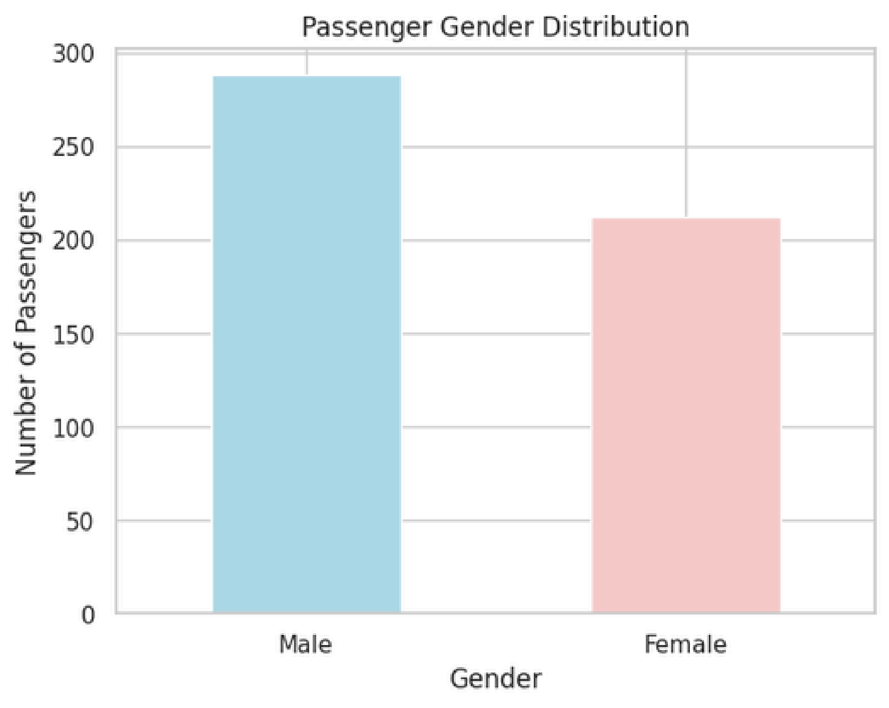
```
survived
1    270
0    230
Name: count, dtype: int64
```



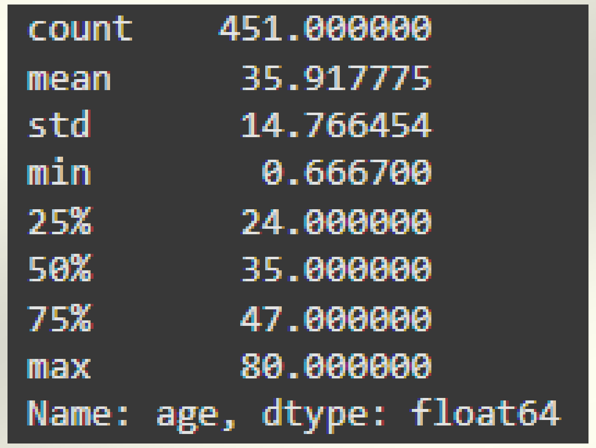
PASSENGER GENDER



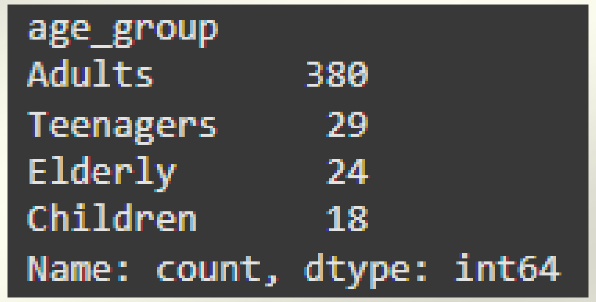
```
sex
male      288
female    212
Name: count, dtype: int64
```



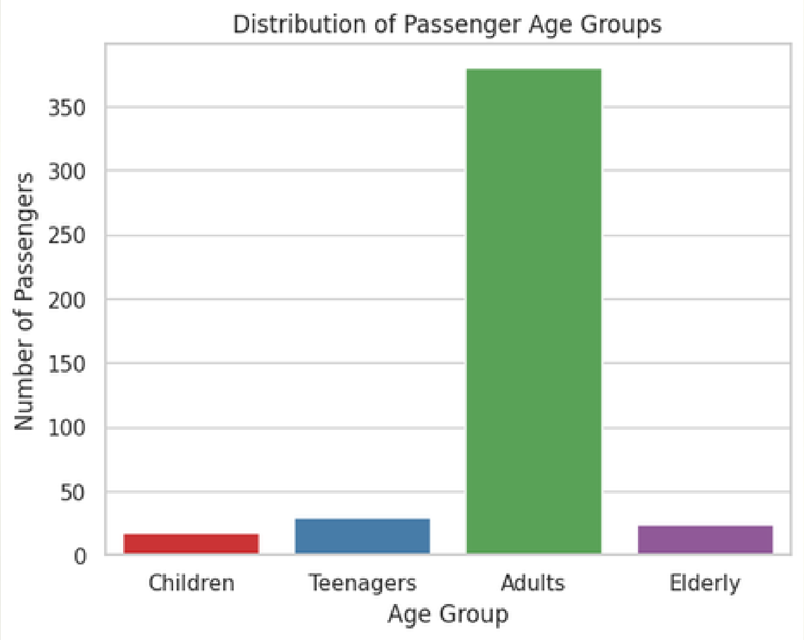
PASSENGER AGE



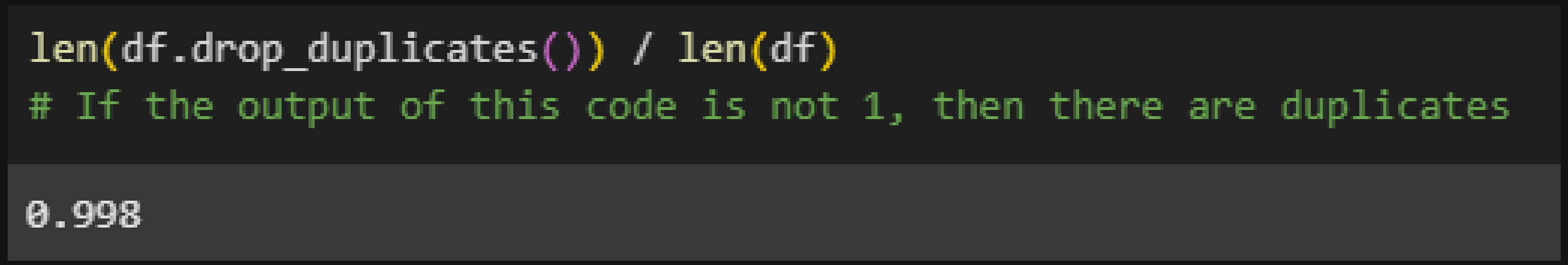
```
count    451.000000
mean      35.917775
std       14.766454
min        0.666700
25%       24.000000
50%       35.000000
75%       47.000000
max       80.000000
Name: age, dtype: float64
```



```
age_group
Adults       380
Teenagers     29
Elderly       24
Children      18
Name: count, dtype: int64
```

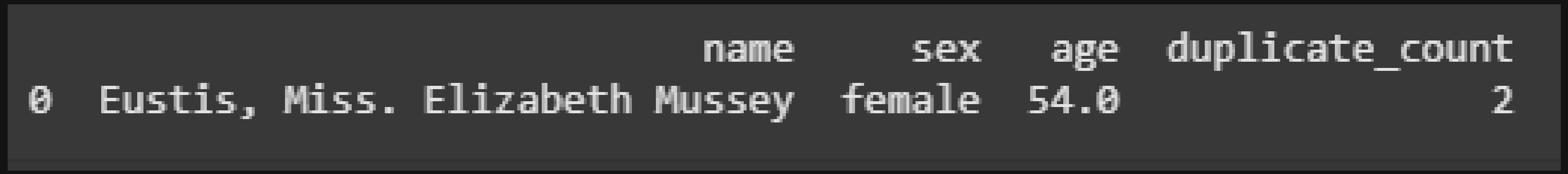


BEFORE



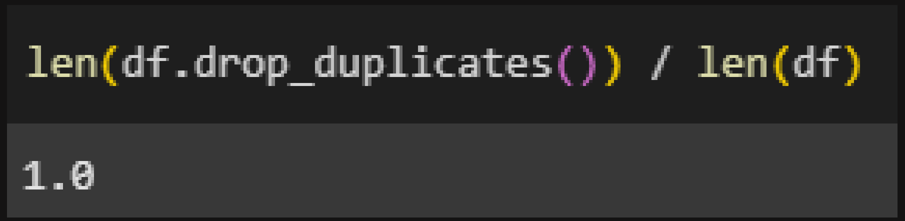
```
len(df.drop_duplicates()) / len(df)
# If the output of this code is not 1, then there are duplicates
```

```
0.998
```



```
                                name     sex   age  duplicate_count
0  Eustis, Miss. Elizabeth Mussey  female  54.0                2
```

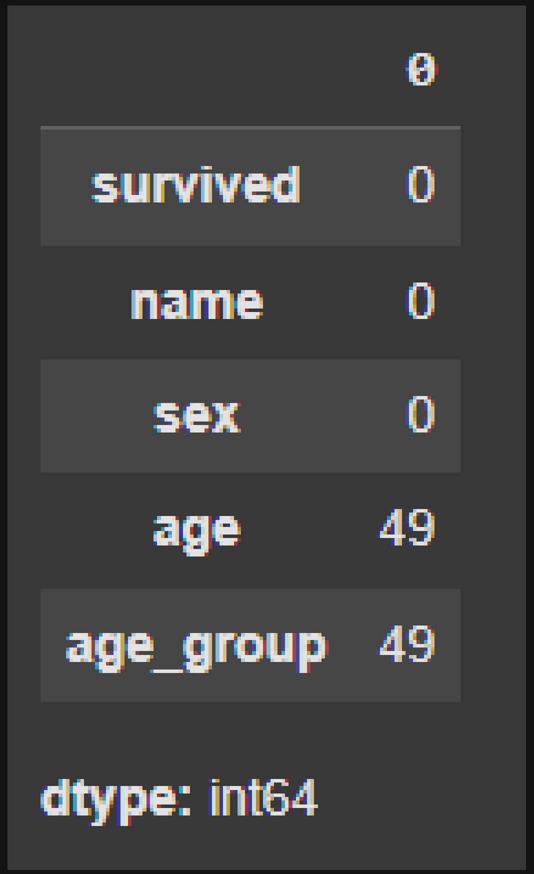
AFTER



```
len(df.drop_duplicates()) / len(df)
```

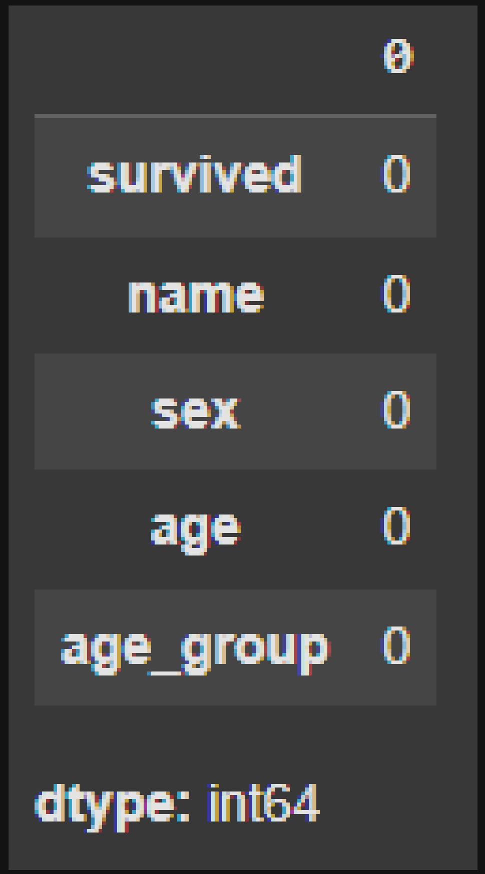
```
1.0
```

BEFORE



```
              0
 survived     0
     name     0
      sex     0
      age    49
age_group    49

dtype: int64
```

AFTER



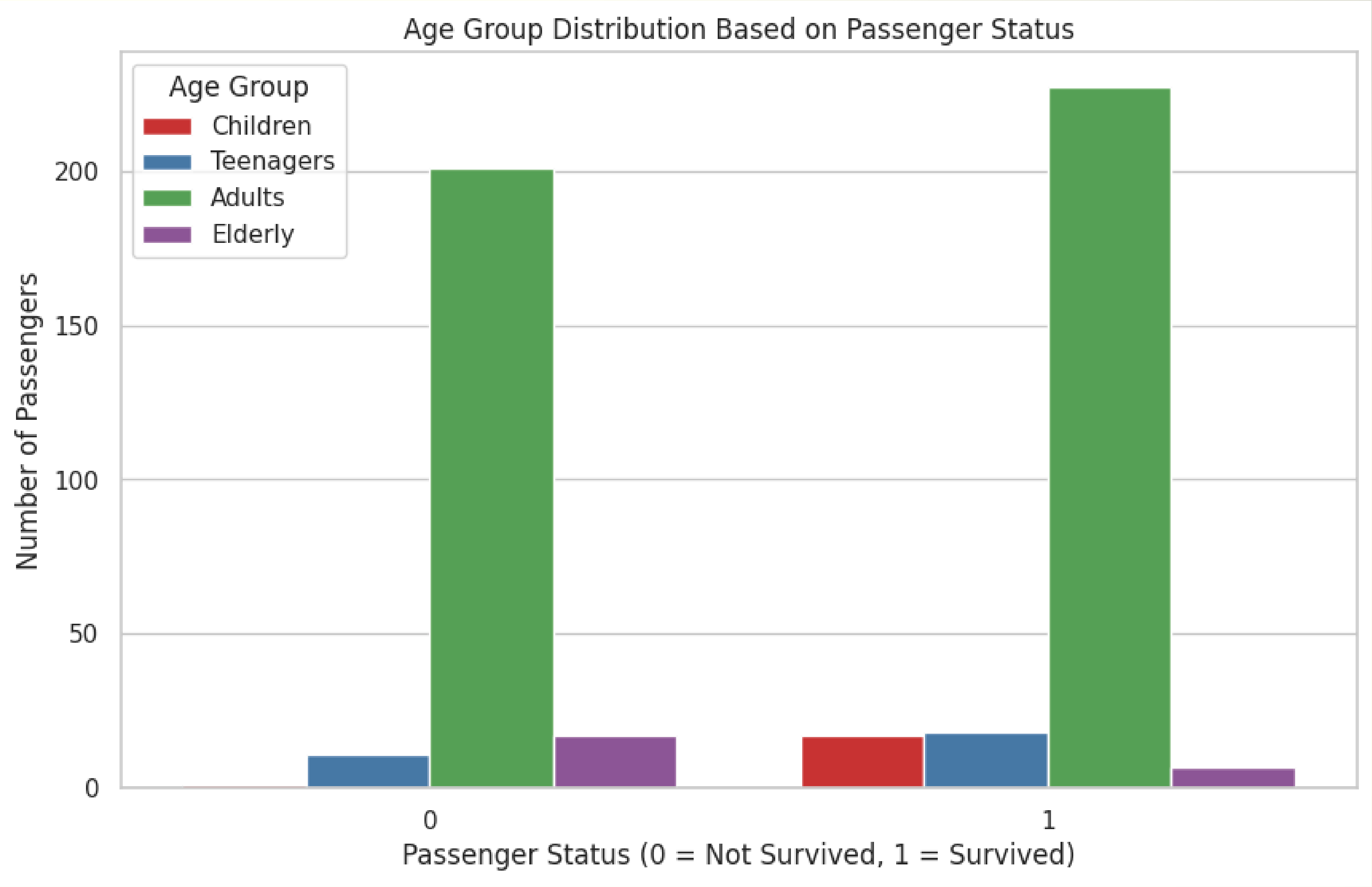
```
              0
 survived     0
     name     0
      sex     0
      age     0
age_group     0

dtype: int64
```

Data Preprocessing

Data Preprocessing is the step to prepare the data before analysis. This phase includes:

- Handling Missing Values: Filling or removing missing values in the data, such as using median for numerical data or mode for categorical data.
- Handling Duplicates: Removing duplicate data to maintain the quality of analysis.

Final Data Visualization and Analysis Results

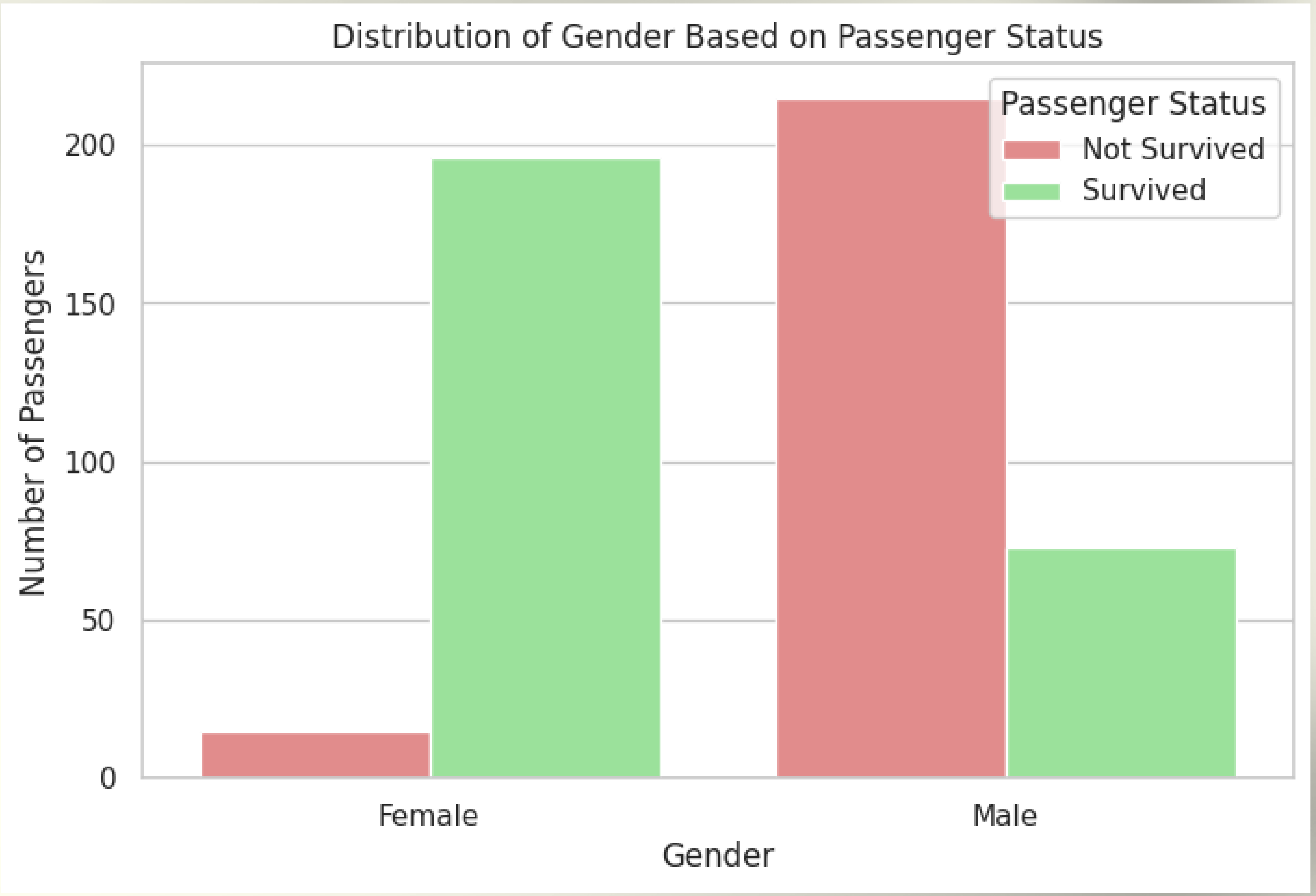


Age Distribution Based on Passenger Status

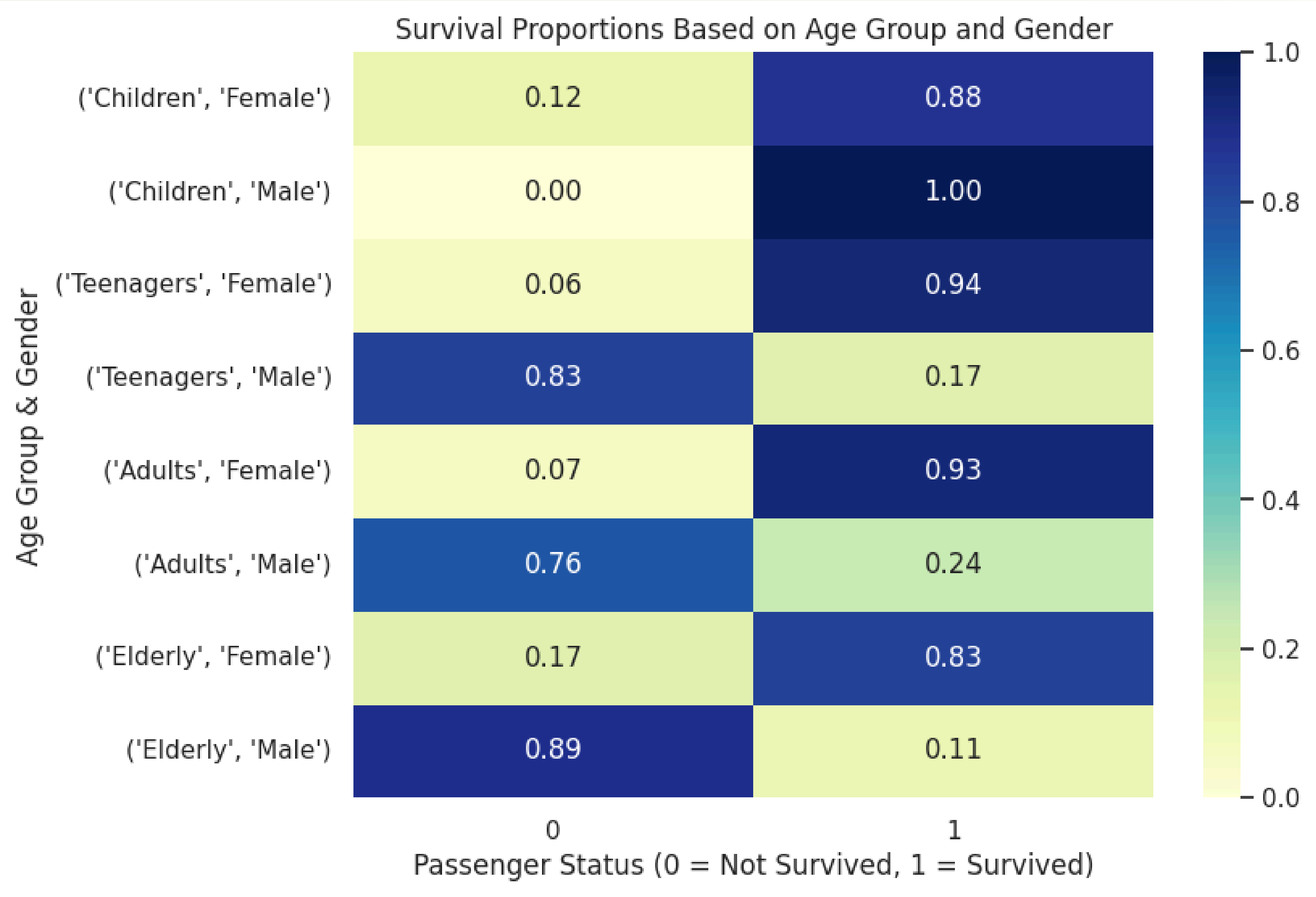
The chart shows that adults make up the largest group, with more survivors than non-survivors. Children and teenagers also have more survivors, though their numbers are smaller. This may suggest a priority for younger age groups during the rescue. The elderly group has more non-survivors, likely due to physical limitations. Overall, age appears to impact survival likelihood.

Gender Distribution Based on Passenger Status

The chart shows that more female passengers survived than male passengers. This may be due to the "women and children first" evacuation policy, which likely gave females a higher chance of survival, while more males did not survive.



Survival Proportions Based on Age Group and Gender



The heatmap shows that females had a higher survival rate than males, especially in the teenage and adult groups. Teenage girls had a 94% survival rate, while teenage boys only had 17%. In the elderly group, elderly males had a much higher survival rate than elderly females. Overall, survival was influenced by both gender and age, with females, especially teenagers and adults, more likely to survive due to priority during evacuation. However, children and the elderly showed different survival patterns.

Thank You