TITANIC PASSENGER SAFETY PREDICTIONS BY AGE AND GENDER

Start Presentation

by : Farrel Farica Firjaturazza

INTRODUCTION

The sinking of the Titanic in 1912 is one of the most tragic and well-known maritime disasters in history. With over 2,200 passengers and crew members on board, only around 700 survived. This tragedy sparked questions that still intrigue us today: What factors influenced someone's chance of survival? Was it their gender, age, social status, or a combination of these?

In this project, I analyze a simplified Titanic dataset provided during the Data Science Bootcamp, which includes information on 500 passengers. Through steps such as data preprocessing, handling missing values, and exploratory data analysis, the goal is to uncover potential patterns behind survival outcomes—particularly focusing on variables like age and gender.

Background

The Titanic disaster remains one of the most talked-about historical tragedies, with over 1,500 lives lost. Many researchers and analysts have explored what factors influenced passenger survival. This project uses a simplified Titanic dataset to investigate survival patterns based on gender and age.

Problem Statement

What demographic factors—such as age and gender—played a role in a passenger's chance of survival during the Titanic disaster?

Urgency

Understanding survival patterns provides not only historical insights but also builds foundational skills in data cleaning, visualization, and analysis—important capabilities for any aspiring data scientist.

Project Goals

The main goal of this project is to perform a structured data analysis using a simplified Titanic dataset. Specifically, this project aims to:

- Understand the distribution of passenger demographics such as age and gender.
- Explore whether certain factors (like age and gender) affected survival rates.
- Practice essential data science skills such as data cleaning, handling missing values, and creating data visualizations.
- Present findings clearly through visual insights and summaries that are easy to interpret.

Data Overview

	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

	survived	name	sex	age
495	1	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.0
496	0	Mangiavacchi, Mr. Serafino Emilio	male	NaN
497	0	Matthews, Mr. William John	male	30.0
498	0	Maybery, Mr. Frank Hubert	male	40.0
499	0	McCrae, Mr. Arthur Gordon	male	32.0

- The dataset consists of 500 rows and 4 columns, where each row represents a unique Titanic passenger. These columns include **survival status, name, gender, and age**. This structured format allows us to analyze demographic characteristics and survival outcomes across different groups.

About Dataset

```
<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
```

The age column has 451 non-null values out of 500, which means there are 49 missing values, while the other 3 columns have no missing values.

Data Preprocessing

- Titanic Dataset -

Import Libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)
```

NumPy: Used for basic math operations such as calculating the average, median, and assisting in the manipulation of numerical data.

Pandas: untuk analisis dan manipulasi data berbentuk tabel (dataframe).

Seaborn: Used to create plots such as bar charts, count plots, and visualization of distributions by category (e.g. gender vs survival rate).

Matplotlib: Used in conjunction with Seaborn to display charts and organize visual aspects such as chart titles, labels, and sizes.

Load Data

```
[ ]  from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ]  # import data
     df = pd.read_excel('/content/drive/MyDrive/CaseProject/titanic.xlsx')
     data = df.copy()
```

- This step aims to connect **Google Colab** with Google Drive, so that the dataset files stored in Drive can be accessed.

- Using the **pd.read_excel()** function from the pandas library to read an Excel (.xlsx) file named titanic.xlsx.

Showing top 5 rows of the data

	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

data.head()

This function is used to display the first 5 rows of the dataset.

	survived	name	sex	age
495	1	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.0
496	0	Mangiavacchi, Mr. Serafino Emilio	male	NaN
497	0	Matthews, Mr. William John	male	30.0
498	0	Maybery, Mr. Frank Hubert	male	40.0
499	0	McCrae, Mr. Arthur Gordon	male	32.0

data.tail()

This function is used to display the last 5 rows of the dataset.

Duplicate Handling

```
[ ]  len(data)

     500

[ ]  len(data.drop_duplicates())

     499

[ ]  len(data.drop_duplicates()) / len(data)

     0.998
```

This line calculates the total number of rows in the data DataFrame. The result shows that there are 500 entries before any cleaning is done.

This line counts the number of unique rows after removing any duplicates. The result is 499, indicating that **one duplicated** row exists in the dataset.

This line calculates the ratio of unique data to total data entries. A result of **0.998** means that 99.8% of the data is unique, with minimal duplication.

Duplicate Handling

```
[ ]  # Take duplicate rows (including the original)
     duplicates = data[data.duplicated(keep=False)]
```

```
[ ]  duplicates
```

	survived	name	sex	age
104	1	Eustis, Miss. Elizabeth Mussey	female	54.0
349	1	Eustis, Miss. Elizabeth Mussey	female	54.0

This line filters out all rows that are duplicates, including the original entries. The keep=False argument ensures all matching rows are returned.

This displays the rows identified as duplicates. In this case, two rows are shown with identical values, confirming the duplication.

After displaying the results, it is clear that a duplicate entry exists between row 104 and row 349, containing identical data. Therefore, one of these rows will be removed to ensure the dataset remains clean and consistent.

Duplicate Handling

```
[ ]  #Handling Drop duplicate
     data = data.drop_duplicates()
```

The line **data = data.drop_duplicates()** removes all duplicate rows from the dataset, keeping only the first occurrence of each unique row. This ensures the data is clean and free from redundancy, which is important for accurate analysis.

```
[ ]  len(data.drop_duplicates()) / len(data)

     1.0
```

The code **len(data.drop_duplicates()) / len(data)** calculates the ratio of unique rows to the total number of rows in the dataset. Since the result is 1, it confirms that there are no duplicate entries remaining—all rows are now unique.

Missing value handling

```
[ ] data.isna().sum()
```

	0
survived	0
name	0
sex	0
age	49

dtype: int64

```
for column in data.columns:
    print(f"============= {column} =================")
    display(data[column].value_counts())
    print()
```

The code data.isna().sum() checks for missing values (NaN) in each column of the dataset and sums them up. From the output, we can see that only the age column has missing values, totaling 49, while the survived, name, and sex columns are complete with no missing entries.

This loop iterates through each column in the dataset and displays the count of unique values for each one. It helps to quickly understand the distribution of data across different columns, which is useful for identifying categorical variables and checking data consistency.

Missing value handling

```
total_rows = len(data)
total_rows
```

```
500
```

This code calculates the total number of rows in the dataset by using len(data) and stores the result in the variable total_rows. It then displays the total number of rows, which helps in understanding the dataset's size.

```
total_rows = len(data)

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 (9.80%)
```

The code analyzes missing values for each column in the dataset. It calculates both the count and percentage of missing entries, then prints the result for each column. From the output, only the age column has missing values—49 entries, which is 9.8% of the total.

The percentage of missing values below 20% so we handle numerically with median, categorical with mode.

Missing value handling

```
[ ]  data['age'].median()
     data['age'].fillna(data['age'].median(), inplace=True)
```

The line data['age'].median() calculates the median of the age column, which represents the middle value when all ages are sorted. This method is often preferred over the mean when handling skewed data or outliers. The next line, data['age'].fillna(data['age'].median(), inplace=True), fills all missing values in the age column with the calculated median. By setting inplace=True, the changes are applied directly to the original DataFrame without needing to assign it to a new variable.

Missing value handling

```
data.isnull().sum()
```

	0
survived	0
name	0
sex	0
age	0

dtype: int64

After filling in the missing values, we will recheck the dataset to ensure that there are no remaining missing values. Additionally, we will verify if any duplicate data entries exist that may need to be addressed.

```
data.info()
```

```
<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       500 non-null    float64
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
```

Based on the provided data information, the dataset is now clean, with no missing values, making it ready for further analysis. It can now be used for the next steps in the analysis process without concerns about missing or incomplete data.

Exploratory Data Analyst

- Titanic Dataset -

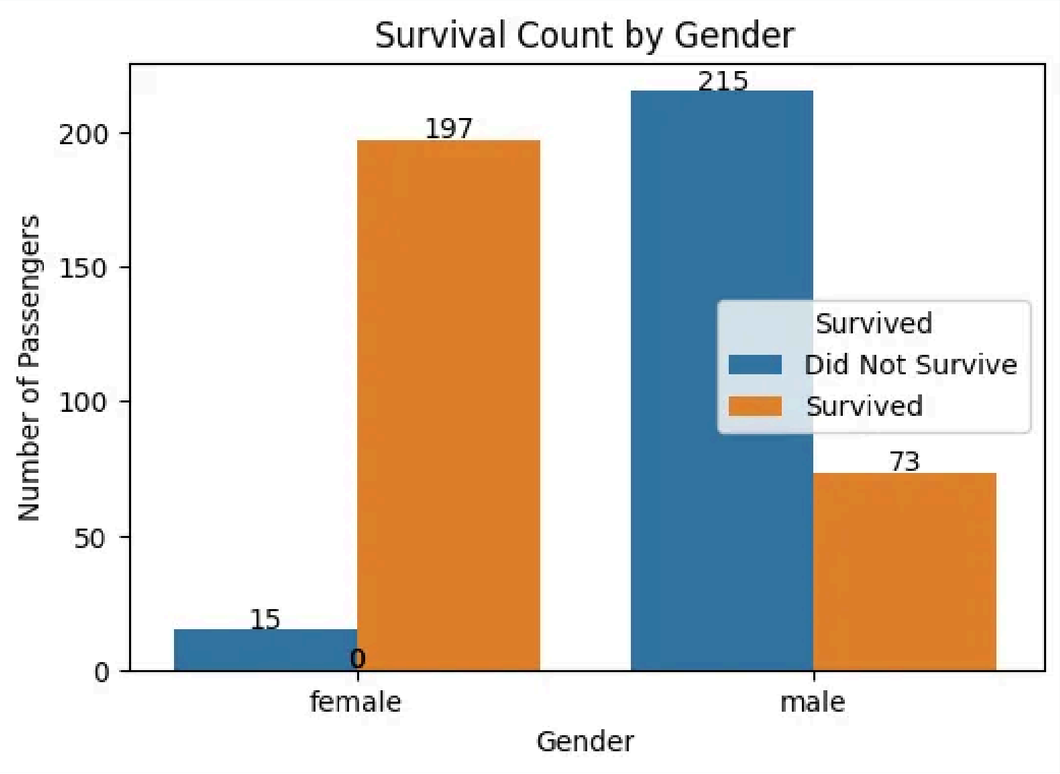
Statistical Summary

	age	survived
count	451.000000	500.000000
mean	35.917775	0.540000
std	14.766454	0.498897
min	0.666700	0.000000
25%	24.000000	0.000000
50%	35.000000	1.000000
75%	47.000000	1.000000
max	80.000000	1.000000

sex	
male	288
female	212

- Age Column:
  - There are 49 empty entries in the age column. This requires reprocessing.
  - The age distribution is quite wide, accounting for 41%.
  - The mean is approximately equal to the median, indicating a fairly symmetrical distribution with minimal skewness.
- Survived Column:
  - The 'survived' column is a binary column, meaning it only contains values of 0 or 1. There's no need to analyze its symmetry. Instead, focus on the balance level (distribution of 0s and 1s).
- Sex Column:
  - The sex field contains 2 unique values: 'male' and 'female'.
  - There are more males, with a total of 288, while the rest are females.
- Duplicate Data:
  - The dataset contains 499 unique values out of 500 entries, indicating one duplicate row.
  - After verifying that one row is exactly the same as another, the duplicate will be removed.

Survival Count by Gender



This bar chart shows the number of passengers who survived and did not survive based on gender. In total, 215 men did not survive and 73 survived, while 15 women did not survive and 197 women survived.

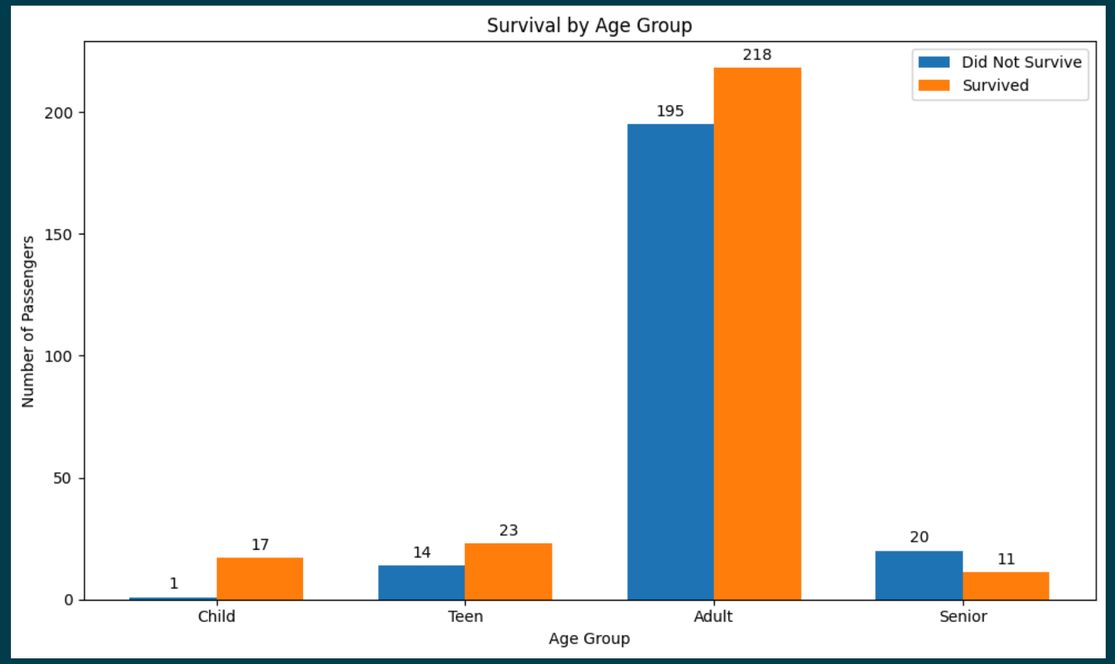
Observations:

- The majority of women (197 out of 212 or about 93%) survived the Titanic disaster.
- The majority of men (215 out of 288 or about 75%) did not survive.
- The total number of passengers in the dataset is 500 (212 women and 288 men).

Indications:

- There was a clear "women and children first" policy in the Titanic evacuation as evident from the data.
- Gender is a very strong factor in determining survival chances.

Survival by Age Group



This bar chart displays the number of passengers who survived and did not survive based on age groups: Child, Teen, Adult, and Senior.

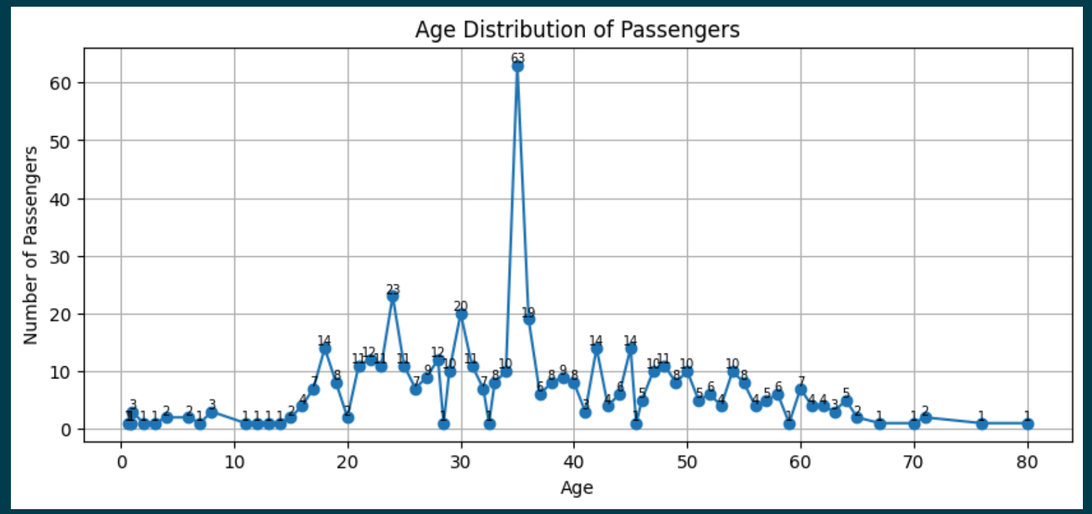
Observations:

- Children had the highest survival rate (17 survived, 1 did not survive).
- Teenagers had a fairly high survival rate (23 survived, 14 did not survive).
- Adults represented the largest group in the data (219 survived, 195 did not survive).
- Seniors had the lowest survival rate (11 survived, 20 did not survive).

Indications:

- Age played an important role in rescue prioritization, in line with the "women and children first" policy.
- Children had the highest rescue priority (about 94% survived).
- Seniors had the lowest survival rate (about 35% survived).

Age Distribution of Passengers



This line graph shows the age distribution of all passengers in the dataset.

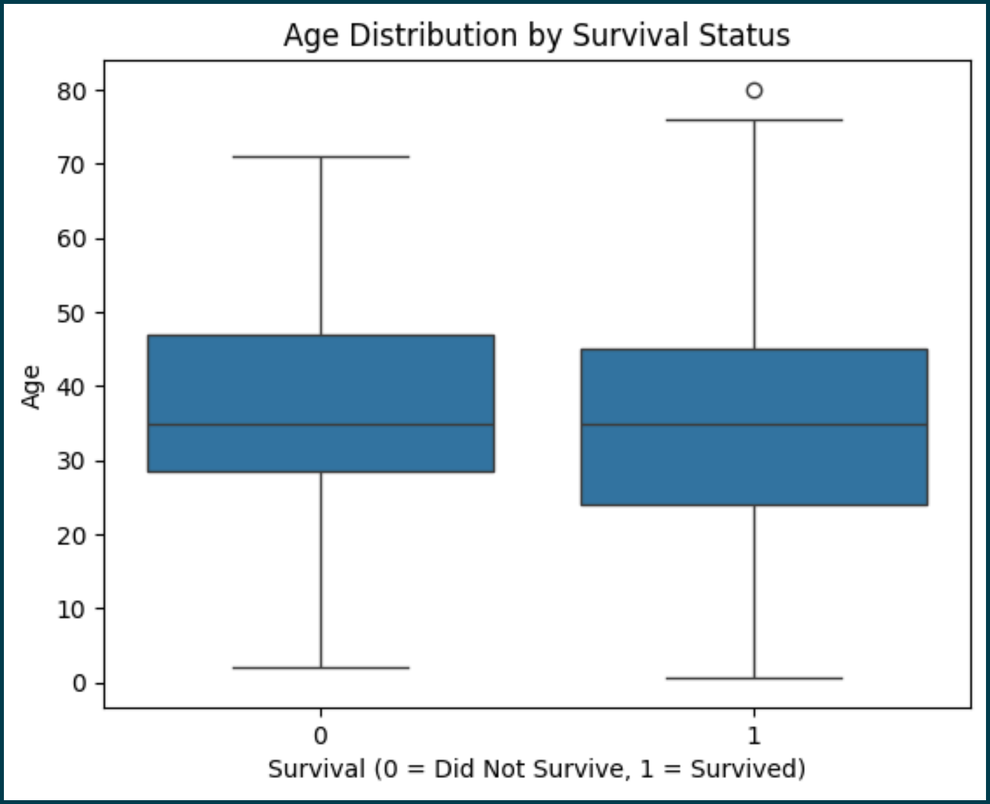
Observations:

- There is a highest peak at around age 34 with more than 60 passengers.
- There are several other peaks around ages 20-35.
- The number of passengers decreases significantly after age 60.
- There are passengers from various age groups, from infants to seniors around 80 years old.

Indications:

- The majority of passengers were adults in their productive age (20-40 years).
- There is an uneven distribution for some ages, particularly age 34 which shows a very high number of passengers.

Age Distribution by Survival Status



This box plot shows the age distribution based on survival status (0 = did not survive, 1 = survived).

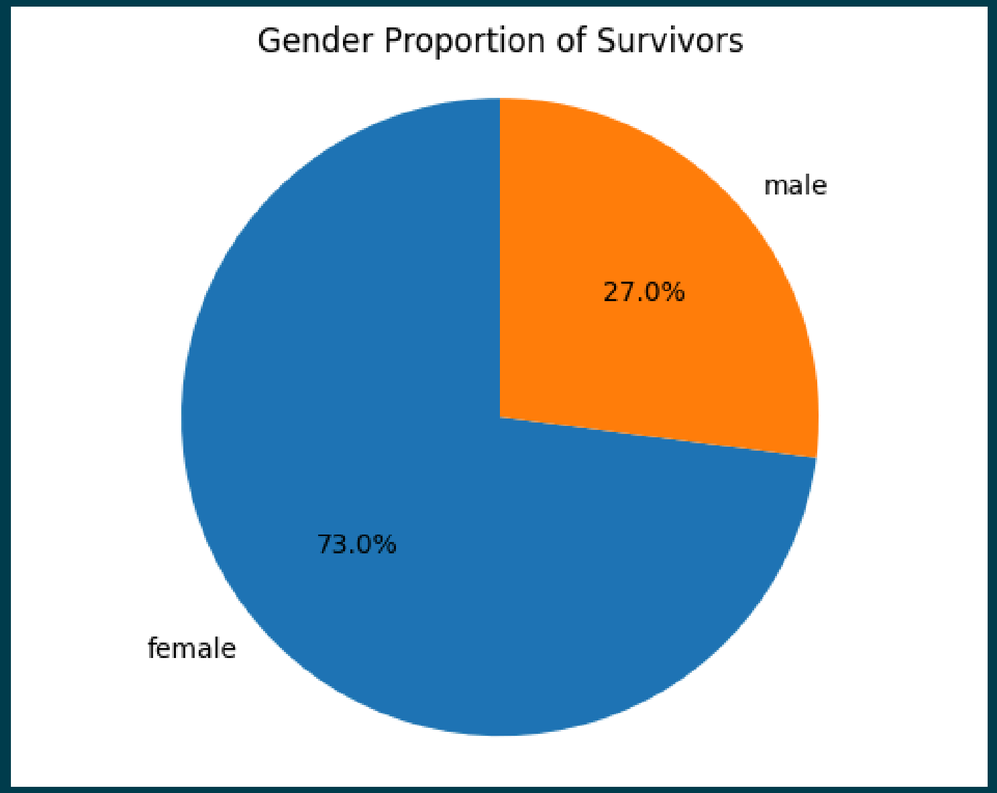
Observations:

- The median age for both groups (survived and did not survive) tends to be similar, around 35 years.
- The age range for both groups is also similar, from about 0-70 years.
- There are some outliers in the survived group (1), indicating some elderly passengers who managed to survive.

Indications:

- In general, the age distribution between those who survived and did not survive does not show a very significant difference.
- Age alone may not be a major determining factor for survival if not combined with other factors such as gender.

Gender Proportion of Survivors



sex	
male	288
female	212

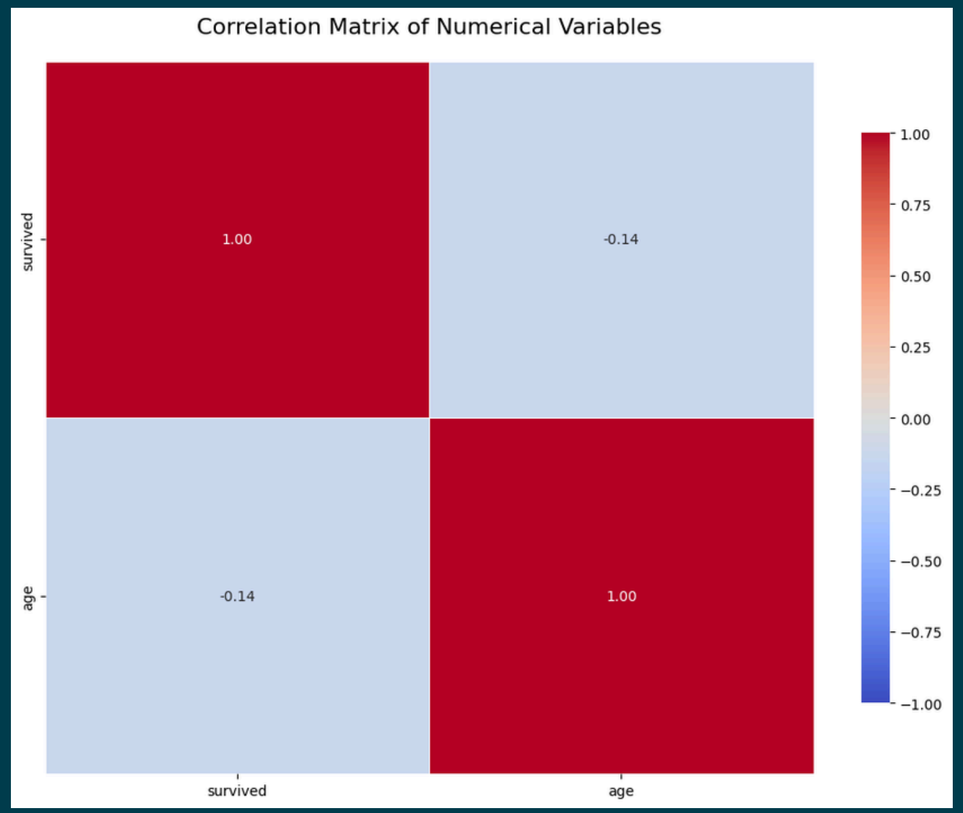
Observations:

- 73% of all survivors were women.
- 27% of all survivors were men.

Indications:

- Women have a much higher representation in the survivor group.
- This data is consistent with the first graph and reinforces the conclusion that gender strongly influenced survival chances.

Correlation Matrix of Numerical Variables



- The diagonal values (survived-survived and age-age) show perfect correlation (1.00) as expected, since any variable is perfectly correlated with itself. These are displayed in dark red.
- There is a weak negative correlation (-0.14) between age and survival. This negative value indicates that as age increases, the likelihood of survival slightly decreases.
- The correlation value is relatively small in magnitude, suggesting that while there is a relationship between age and survival, it is not particularly strong on its own.
- The correlation matrix is symmetrical, with the same values appearing in both the upper and lower triangles (-0.14 appears twice).
- The color scale ranges from dark blue (strong negative correlation) through light blue (weak negative correlation), white (no correlation), and various shades of red (positive correlation), with dark red indicating strong positive correlation.

Conclusion

- Titanic Dataset -

1. Survived

The survived column has a strong relationship with gender but a weak relationship with age (correlation -0.14). The data shows that 270 out of 500 passengers (54%) survived, with a very uneven distribution between genders.

2. Sex

The visualization shows the highly significant effect of gender on survival:

- Women had a very high survival rate (93% survived)
- Men had a low survival rate (only 25% survived)
- Of the passengers who survived, 73% were women and 27% were men
- Gender comparison in the dataset: 212 women (42%) and 288 men (58%)

3. Name

Name can be used to extract additional information such as titles (Mr., Mrs., Miss, etc.) which can provide further insight into social or marital status.

4. Age

- There is an uneven pattern of age distribution with a significant peak around age 34
- Passengers range from infants to 80 year olds
- Passenger age distribution tends to be normal with most being in the productive age (20-40 years)
- Weak negative correlation (-0.14) with survival
- The relationship with survival is clearer when grouped:
  - Children (0-12): 94% safety rate
  - Adolescents (13-19): 62% safety rate
  - Adults (20-59): 53% safety rate
  - Elderly (60+): 35% survival rate

Thank You

- Titanic Dataset -