Exploratory Data Analysis on Titanic Dataset

Dataset source: Titanic.csv
Number of records: 891

Objective

- To analyze the factors that influenced passenger survival during the Titanic .
- The goal is to identify key patterns, relationships, and insights that explain survival outcomes.

Table of Contents

1. Introduction

Objective: Explore the Titanic dataset to find patterns that affected passenger survival. The report includes data preview, cleaning steps, summary statistics, visualizations, and insights.

2. Data Loading & Preview

First 10 rows of the original dataset:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	nan	S
2	1	1	Cumings, Mrs. John Bradley	female	38.0	1	0	PC 17599	71.2833	C85	C

			(Florence Briggs Thayer)								
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	nan	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05	nan	S
6	0	3	Moran, Mr. James	male	nan	0	0	330877	8.4583	nan	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075	nan	S
9	1	3	Johnson, Mrs. Oscar W (Elisabe	female	27.0	0	2	347742	11.1333	nan	S

			th Vilhelmina Berg)								
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	nan	C

3. Missing Values & Cleaning

Missing values before cleaning:

Column	MissingCount	MissingPercent
PassengerId	0	0.0
Survived	0	0.0
Pclass	0	0.0
Name	0	0.0
Sex	0	0.0
Age	177	19.87
SibSp	0	0.0
Parch	0	0.0
Ticket	0	0.0
Fare	0	0.0
Cabin	687	77.1
Embarked	2	0.22

Cleaning steps applied:

- Filled Age with median.

- Filled Embarked with mode.
- Dropped Cabin column due to many missing values (if present).

4. Feature Engineering

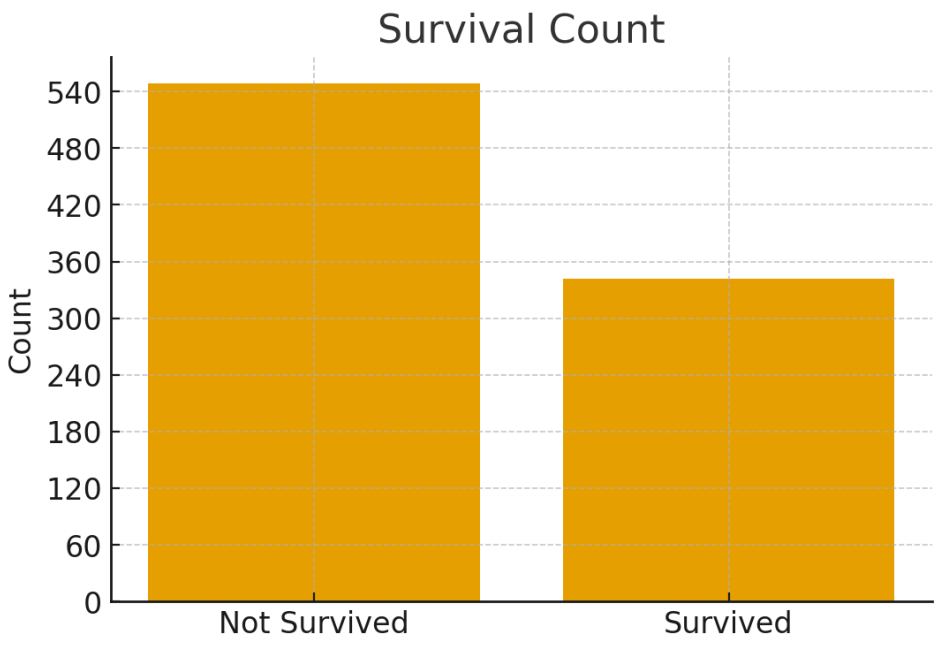
Added a 'FamilySize' column (SibSp + Parch + 1) and a 'FamilyBucket' column (Alone / Small / Large).

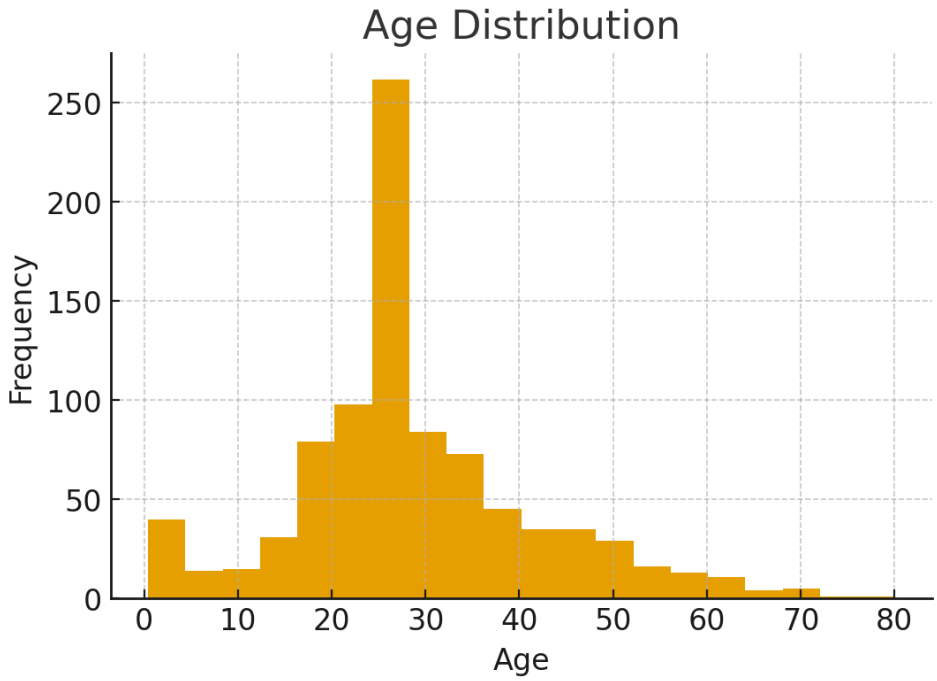
Survival rate by family bucket:

FamilyBucket	SurvivalRate
Alone	0.304
Large	0.161
Small	0.579

5. Univariate Analysis

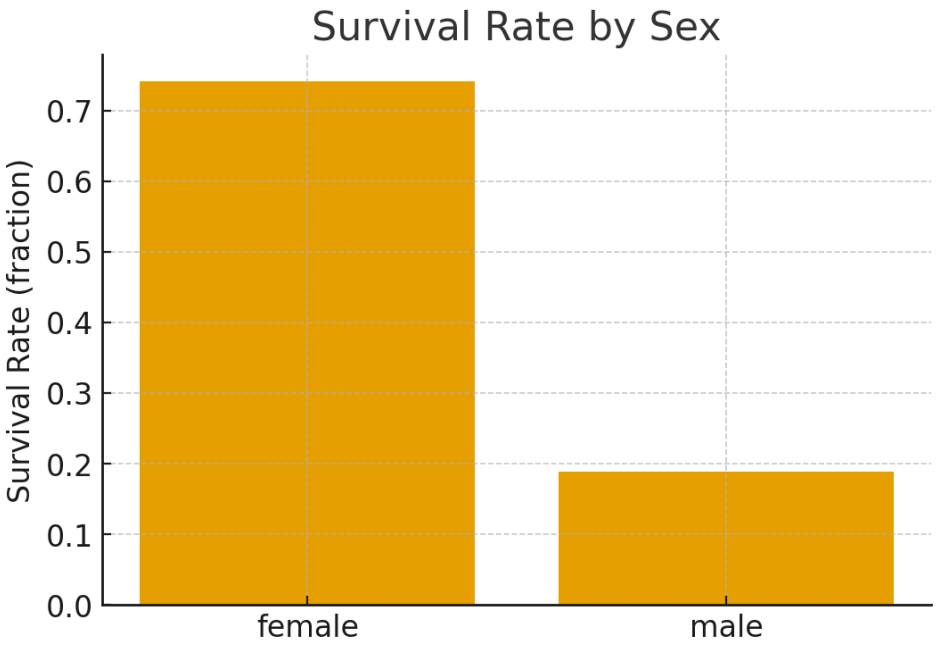
Key univariate visuals: Survival counts and Age distribution.

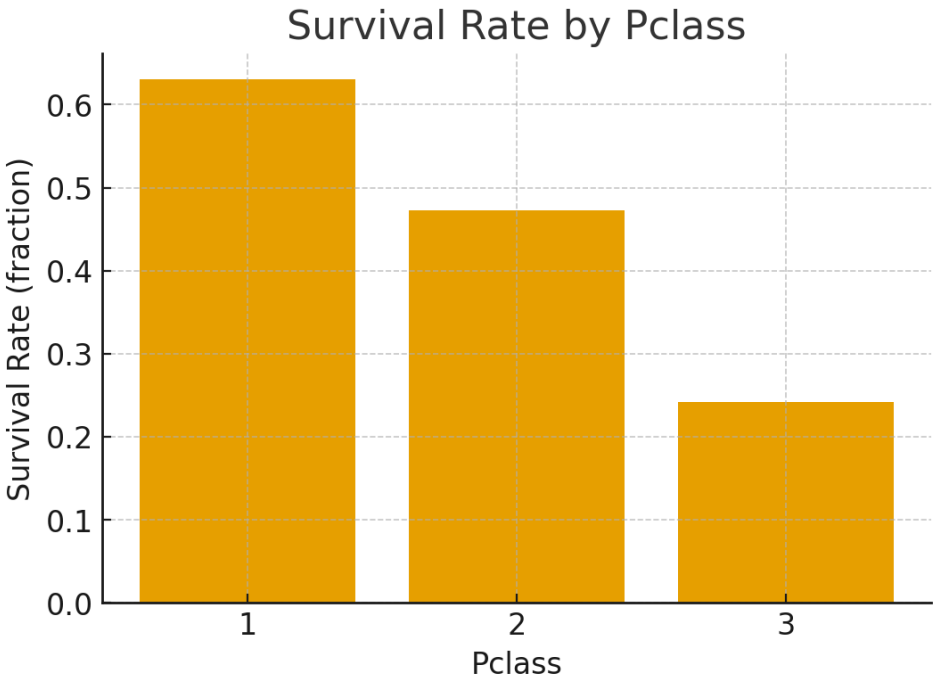


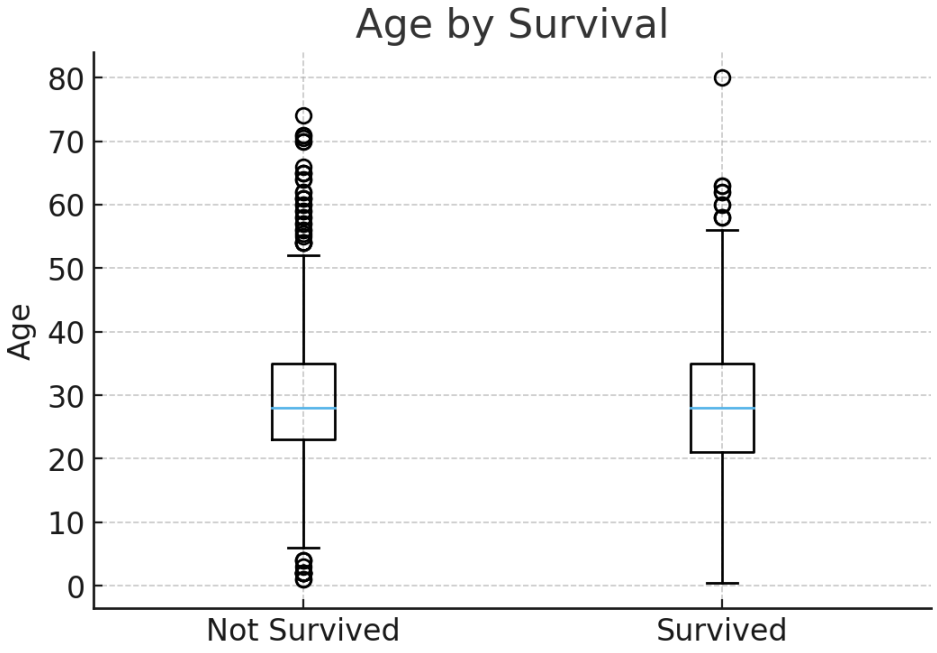


6. Bivariate Analysis

Survival rate by Sex and by Pclass, and Age distribution by survival.

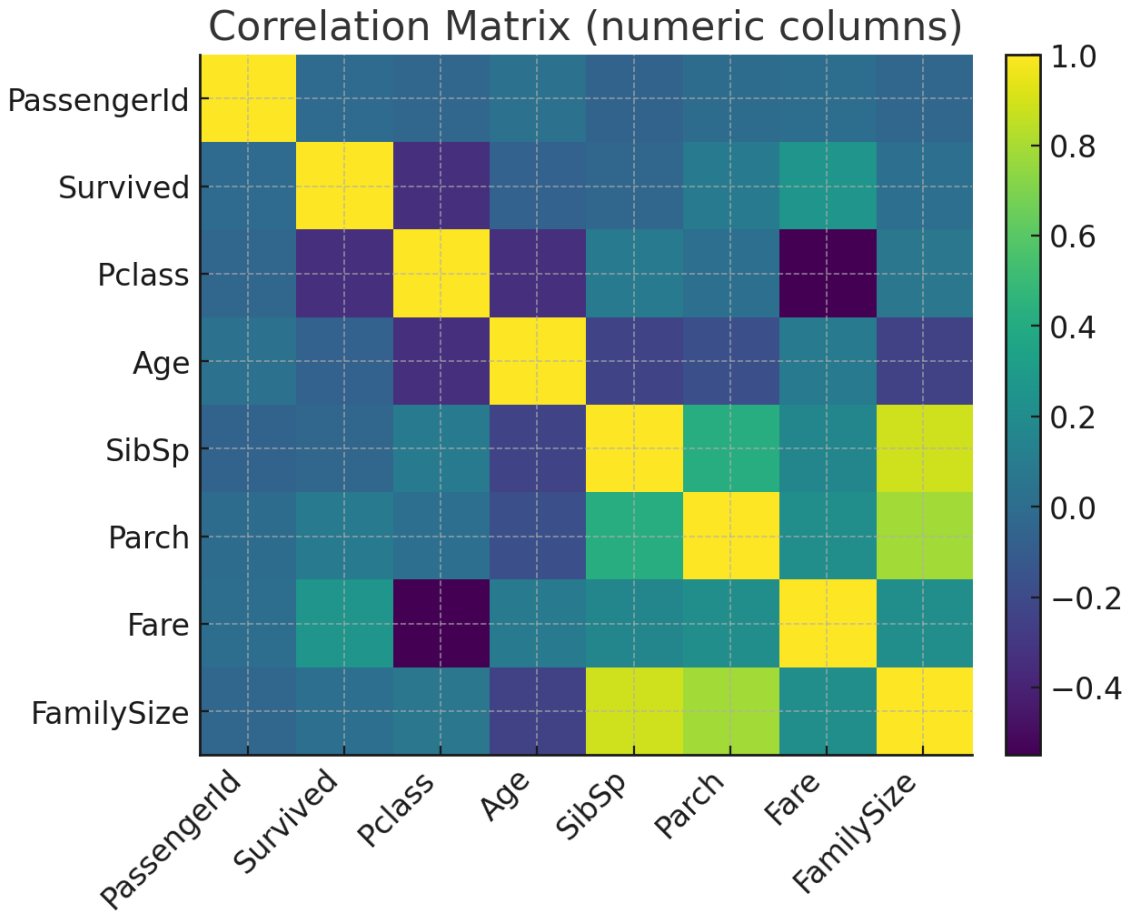



Survival Rate by Pclass
Survival Rate (fraction)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
Pclass


Age by Survival
Age
0
10
20
30
40
50
60
70
80
Not Survived
Survived

7. Correlation Analysis

Correlation matrix of numeric columns:



Correlation (numeric) summary:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	FamilySize
PassengerId	1.000	-0.005	-0.035	0.034	-0.058	-0.002	0.013	-0.040
Survived	-0.005	1.000	-0.338	-0.065	-0.035	0.082	0.257	0.017
Pclass	-0.035	-0.338	1.000	-0.340	0.083	0.018	-0.549	0.066
Age	0.034	-0.065	-0.340	1.000	-0.233	-0.172	0.097	-0.246
SibSp	-0.058	-0.035	0.083	-0.233	1.000	0.415	0.160	0.891

Parch	-0.002	0.082	0.018	-0.172	0.415	1.000	0.216	0.783
Fare	0.013	0.257	-0.549	0.097	0.160	0.216	1.000	0.217
FamilySize	-0.040	0.017	0.066	-0.246	0.891	0.783	0.217	1.000

8. Insights & Conclusions

- Women had a higher survival rate compared to men.
- First class (Pclass=1) passengers had higher survival rates.
- Family size influenced survival: passengers travelling alone had different survival rates compared to small/large families.
- Age distribution shows a concentration of passengers in young to middle ages; age influenced survival to some extent.

9. Code Appendix

Below are the main code snippets used to generate the analysis and visuals.

Loading & basic info

```
import pandas as pd
df = pd.read_csv("Titanic.csv")
df.info()
df.head()
```

Cleaning

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(columns=['Cabin'], inplace=True)
```

Feature engineering

```
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
def fam_bucket(n):
    if n == 1:
        return 'Alone'
    elif 2 <= n <= 4:
        return 'Small'
    else:
        return 'Large'
```

```
df['FamilyBucket'] = df['FamilySize'].apply(fam_bucket)
```

plot (survival count)

```
import matplotlib.pyplot as plt
counts = df['Survived'].value_counts().sort_index()
plt.bar(['Not Survived','Survived'], counts.values)
plt.title('Survival Count')
plt.ylabel('Count')
plt.show()
```