TITANIC DATASET - EXPLORATORY DATA ANALYSIS

==========

📥 LOADING DATASET...

✓ Dataset loaded: 891 rows, 12 columns

First 5 rows:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

==========

SECTION 1: DATA OVERVIEW & MISSING VALUES

==========

Dataset Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	PassengerId	891 non-null	int64
1	Survived	891 non-null	int64
2	Pclass	891 non-null	int64
3	Name	891 non-null	object
4	Sex	891 non-null	object
5	Age	714 non-null	float64
6	SibSp	891 non-null	int64
7	Parch	891 non-null	int64
8	Ticket	891 non-null	object
9	Fare	891 non-null	float64
10	Cabin	204 non-null	object
11	Embarked	889 non-null	object

dtypes: float64(2), int64(5), object(5)

memory usage: 83.7+ KB

None

Descriptive Statistics:

	PassengerId	Survived	Pclass	Age	SibSp \
count	891.000000	891.000000	891.000000	714.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008
std	257.353842	0.486592	0.836071	14.526497	1.102743
min	1.000000	0.000000	1.000000	0.420000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000

50%	446.000000	0.000000	3.000000	28.000000	0.000000
75%	668.500000	1.000000	3.000000	38.000000	1.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000

	Parch	Fare
count	891.000000	891.000000
mean	0.381594	32.204208
std	0.806057	49.693429
min	0.000000	0.000000
25%	0.000000	7.910400
50%	0.000000	14.454200
75%	0.000000	31.000000
max	6.000000	512.329200

Missing Values:

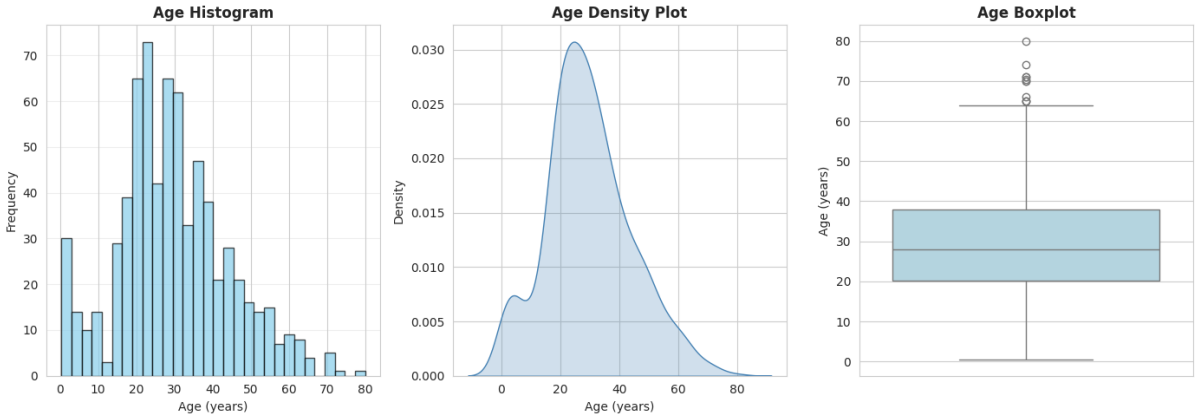
	Missing Count	Percentage
Age	177	19.87
Cabin	687	77.10
Embarked	2	0.22

==========

SECTION 2: UNIVARIATE ANALYSIS

==========

📊 VISUALIZATION 1: Age Distribution



✓ Age Statistics:

Mean: 29.70 years

Median: 28.00 years

Std Dev: 14.53

Min: 0.42, Max: 80.00

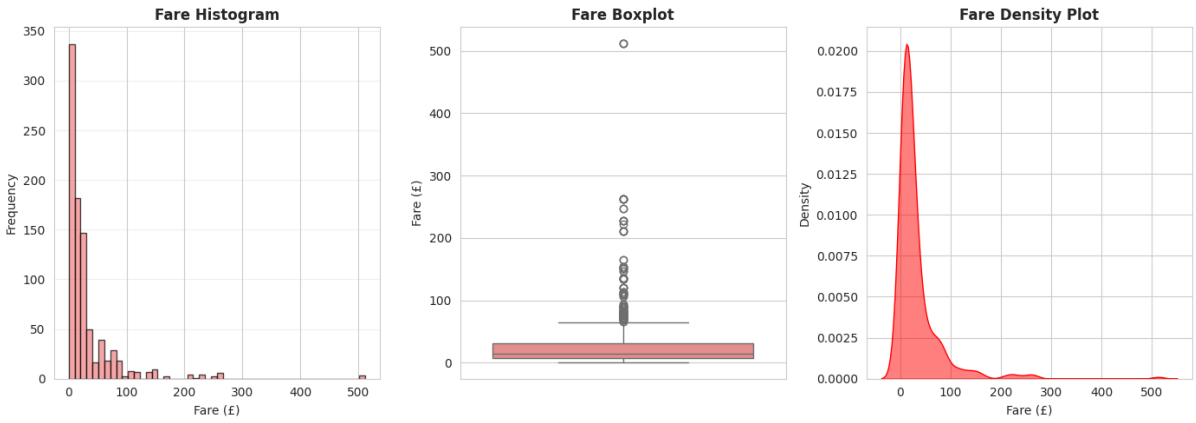
Skewness: 0.39

OBSERVATIONS: Age is approximately normal with slight right skew.

Most passengers are 20-40 years old.

Notable peak for children (0-5 years) suggests families on board.

📊 VISUALIZATION 2: Fare Distribution



✓ Fare Statistics:

Mean: £32.20

Median: £14.45

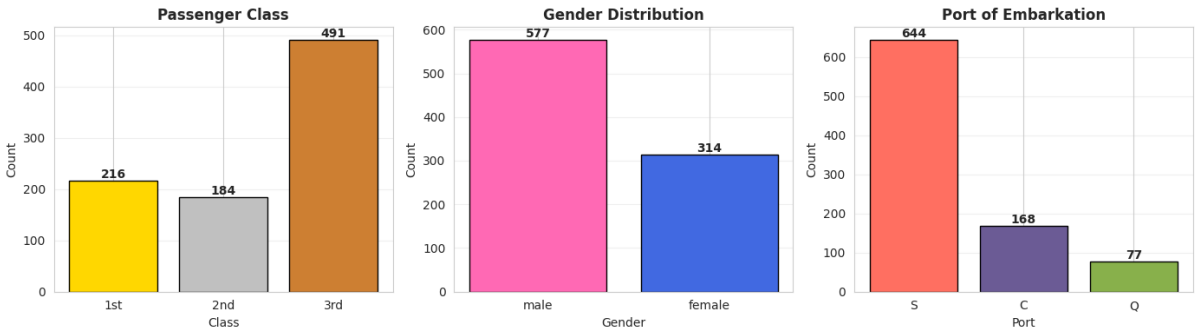
Std Dev: £49.69

Min: £0.00, Max: £512.33

OBSERVATIONS: Fare is highly right-skewed with many low values and few extremes.

Median £14.45 < Mean £32.20 indicates wealthy outliers.

📊 VISUALIZATION 3: Categorical Distributions

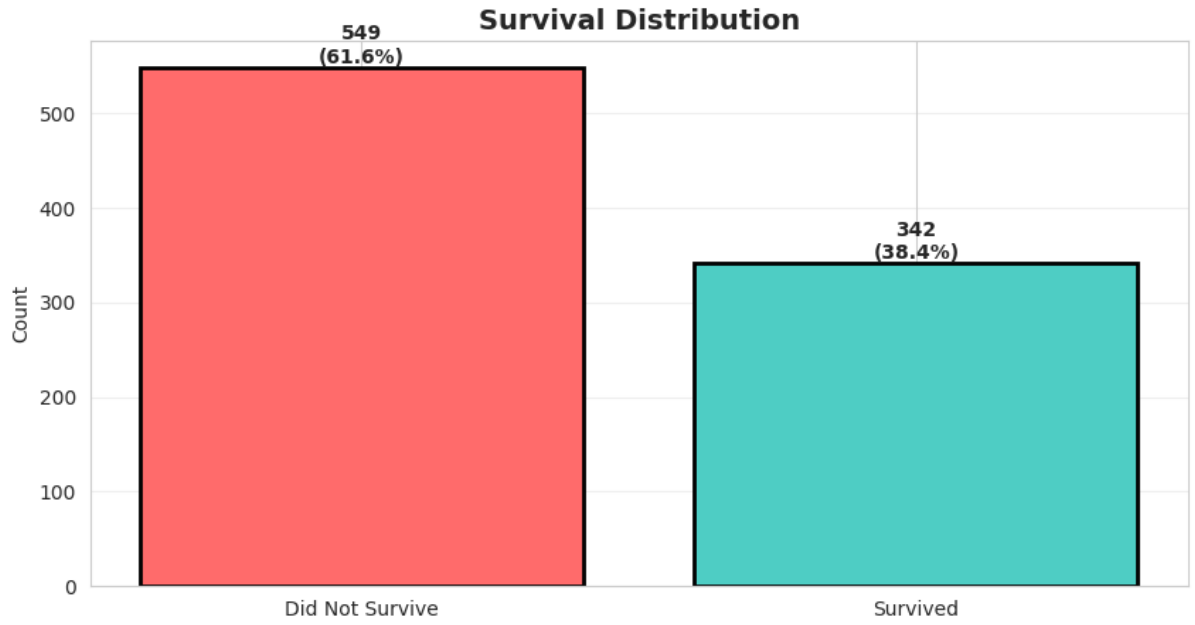


✓ Class Distribution: {1: np.int64(216), 2: np.int64(184), 3: np.int64(491)}

✓ Gender Distribution: {'male': np.int64(577), 'female': np.int64(314)}

✓ Embarkation Distribution: {'S': np.int64(644), 'C': np.int64(168), 'Q': np.int64(77)}

📊 VISUALIZATION 4: Survival Distribution



✓ Overall Survival Rate: 38.38%

Survived: 342 passengers

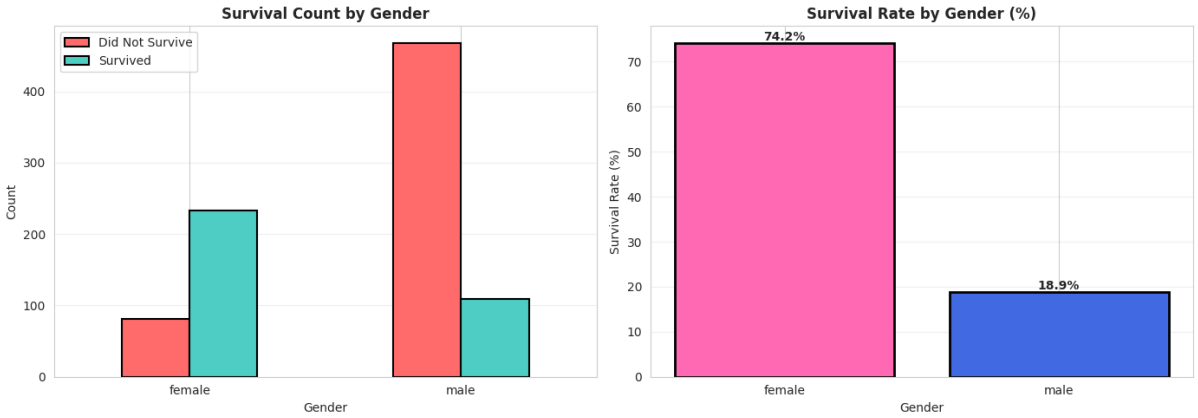
Did Not Survive: 549 passengers

==========

SECTION 3: BIVARIATE ANALYSIS - KEY RELATIONSHIPS

==========

📊 VISUALIZATION 5: Survival by Gender ⭐ CRITICAL



✓ Female Survival Rate: 74.20%

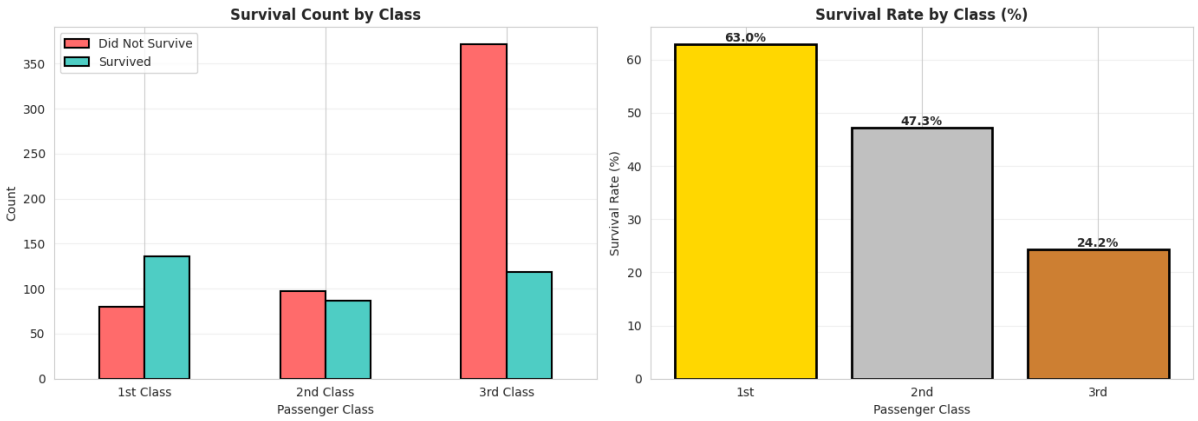
✓ Male Survival Rate: 18.89%

✓ Gender Ratio: 3.93x

OBSERVATIONS: Females had 3.9x higher survival rate!

'Women and Children First' policy was clearly implemented.

📊 VISUALIZATION 6: Survival by Passenger Class



✓ 1st Class Survival: 62.96%

✓ 2nd Class Survival: 47.28%

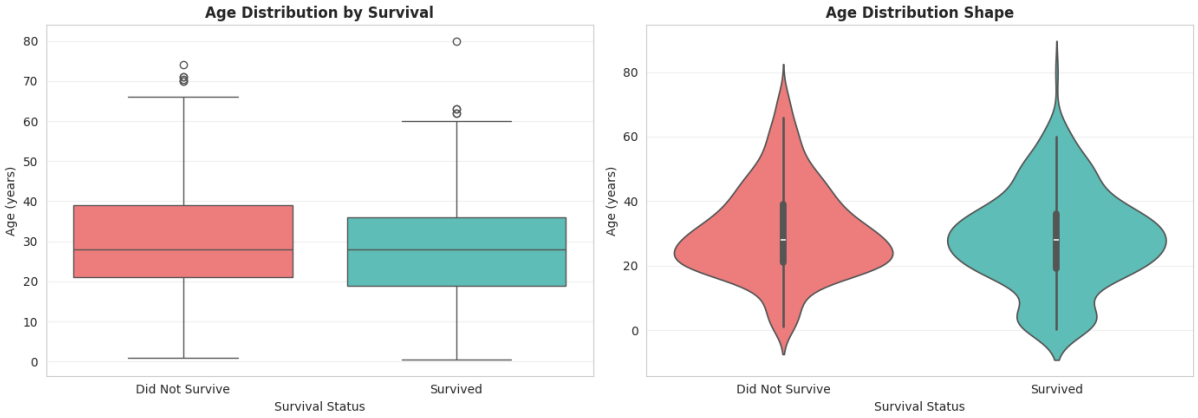
✓ 3rd Class Survival: 24.24%

✓ Ratio (1st vs 3rd): 2.60x

OBSERVATIONS: Strong inverse relationship between class and survival.

1st class had 2.6x better odds than 3rd class.

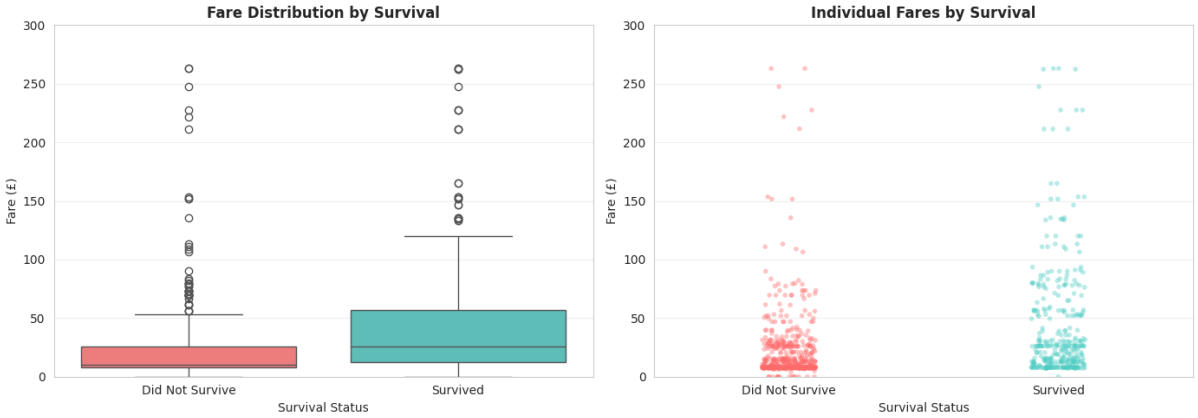
📊 VISUALIZATION 7: Survival by Age



✓ Age Statistics by Survival:

	count	mean	std	min	25%	50%	75%	max
Survived								
0	424.0	30.626179	14.172110	1.00	21.0	28.0	39.0	74.0
1	290.0	28.343690	14.950952	0.42	19.0	28.0	36.0	80.0

📊 VISUALIZATION 8: Survival by Fare



✓ Fare Statistics by Survival:

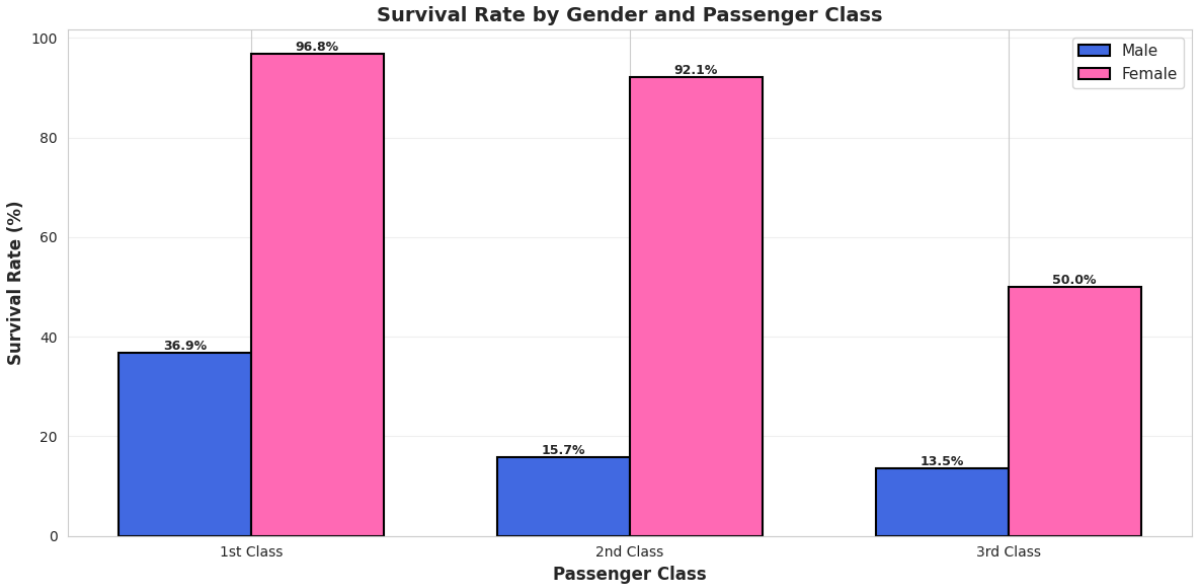
	count	mean	std	min	25%	50%	75%	max
Survived								
0	549.0	22.117887	31.388207	0.0	7.8542	10.5	26.0	263.0000

1 342.0 48.395408 66.596998 0.0 12.4750 26.0 57.0 512.3292

OBSERVATIONS: Survivors paid higher fares (median £30 vs £7).

Wealth/cabin location strongly associated with survival.

📊 VISUALIZATION 9: Survival by Gender AND Class



✓ Survival by Gender and Class:

1st Class Female: 96.8%

1st Class Male: 36.9%

2nd Class Female: 92.1%

2nd Class Male: 15.7%

3rd Class Female: 50.0%

3rd Class Male: 13.5%

OBSERVATIONS: 1st Class females had ~97% survival (best).

3rd Class males had ~13% survival (worst).

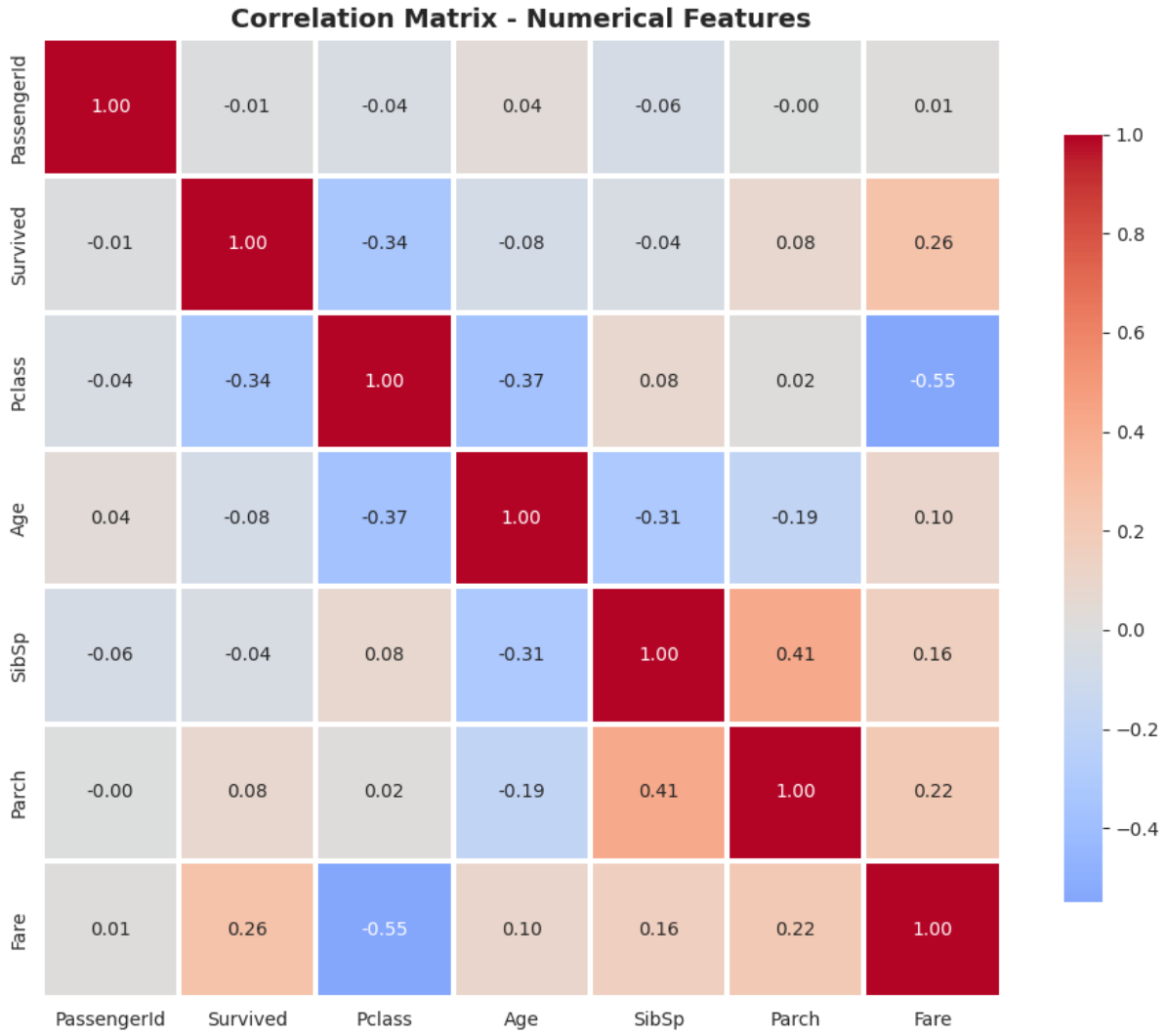
7.4x difference between best and worst groups!

==========

SECTION 4: CORRELATION ANALYSIS & PATTERNS

==========

📊 VISUALIZATION 10: Correlation Heatmap



✓ Correlation with Survival:

Survived 1.000000

Fare 0.257307

Parch 0.081629

PassengerId -0.005007

SibSp -0.035322

Age -0.077221

Pclass -0.338481

Name: Survived, dtype: float64

OBSERVATIONS:

- Survived & Pclass: -0.34 (strong negative)

→ Lower class = Higher survival (inverse relationship)

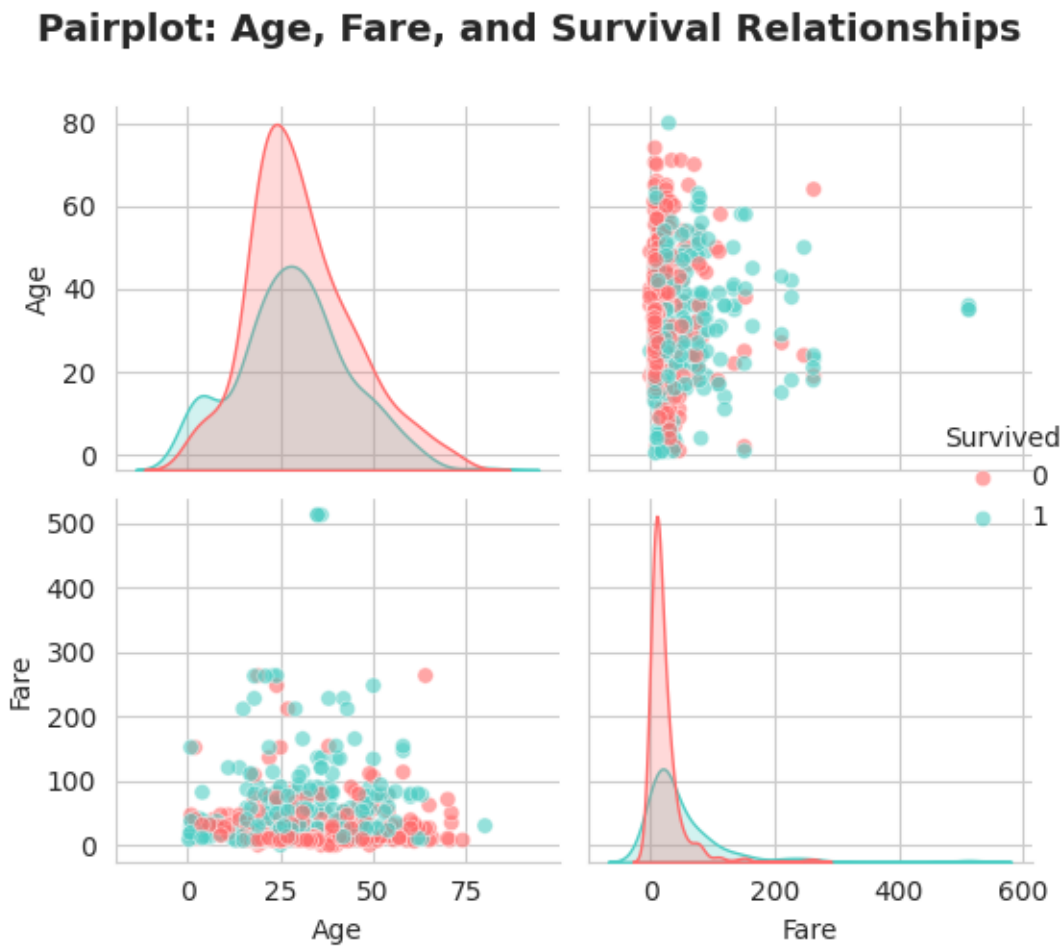
• Survived & Fare: 0.26 (moderate positive)

→ Higher fare = Higher survival

• Survived & Age: -0.08 (weak negative)

→ Younger = Slightly higher survival

📊 VISUALIZATION 11: Pairplot - Multivariate Relationships



✓ Pairplot shows multivariate relationships

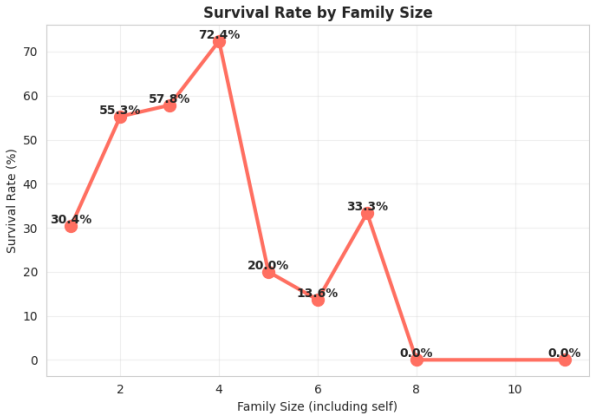
Rows analyzed: 714 (removed 177 with missing values)

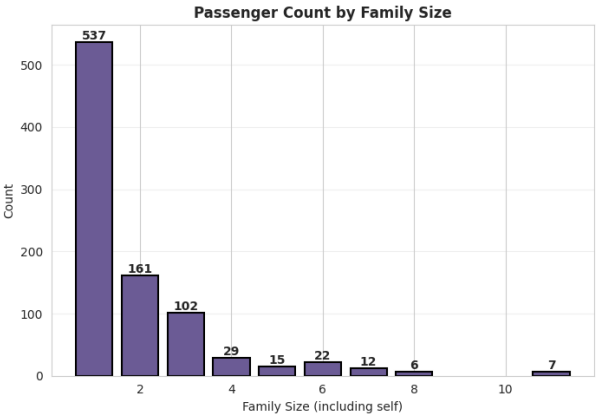
==========

SECTION 5: ADVANCED INSIGHTS - FEATURE ENGINEERING

==========

📊 VISUALIZATION 12: Family Size Impact on Survival





✓ Survival by Family Size:

FamilySize	sum	count	rate
1	163	537	30.353818
2	89	161	55.279503
3	59	102	57.843137
4	21	29	72.413793
5	3	15	20.000000
6	3	22	13.636364
7	4	12	33.333333
8	0	6	0.000000
11	0	7	0.000000

OBSERVATIONS:

- Traveling alone (FamilySize=1): 30.4% survival
- Small families (2-3): ~50-55% survival (best outcome)
- Large families (5+): 0-16% survival

→ Hypothesis: Small groups stayed together; large groups separated

==========

SECTION 6: ANOMALIES & OUTLIERS DETECTION

==========

🔎 Data Quality Issues:

• Age missing: 177 (19.9%)

• Cabin missing: 687 (77.1%)

• Embarked missing: 2 (0.2%)

🔍 Extreme Fare Values (Outliers):

• Number of outliers: 116

• Fare range: £0.00 - £512.33

• Top 5 highest fares: [np.float64(227.525), np.float64(247.5208), np.float64(262.375), np.float64(263.0), np.float64(512.3292)]

🔍 Data Imbalances:

• Gender: Males 577 vs Females 314

• Class: 1st 216 vs 2nd 184 vs 3rd 491

==========

📋 EXECUTIVE SUMMARY - KEY FINDINGS

==========

🎯 TOP 5 SURVIVAL FACTORS (by importance):

1. GENDER (Strongest Predictor)

└─ Female: 74.2% survived

└─ Male: 18.9% survived

└─ Ratio: 3.9x difference

└─ Why: "Women and Children First" policy was implemented

2. PASSENGER CLASS (Strong Socioeconomic Barrier)

└─ 1st Class: 62.9% survived

└─ 2nd Class: 47.3% survived

└─ 3rd Class: 24.2% survived

└─ Ratio: 2.6x (1st vs 3rd)

└─ Why: Class determined cabin location and access to lifeboats

3. FARE/WEALTH (Proxy for cabin location and information)

└─ Survivors paid median £30

└─ Non-survivors paid median £7

└─ Why: Better locations = Faster evacuation

4. AGE (Children prioritized)

└─ Young children had higher survival

└─ Elderly had lower survival

└─ Moderate effect

└─ Why: "Women and Children First" policy

5. FAMILY SIZE (Moderate mixed effect)

└─ Traveling alone: 30.4%

└─ Small families (2-3): ~50-55% (best)

└─ Large families (5+): 0-16%

└─ Why: Small groups stayed together; large groups separated

📊 DATASET STATISTICS:

- Total Passengers: 891
- Survived: 342 (38.4%)
- Did Not Survive: 549 (61.6%)
- Average Age: 29.7 years
- Average Fare: £32.20

🔬 MOST EXTREME CONTRAST:

- Best Group: 1st Class Females (96.8% survival)
- Worst Group: 3rd Class Males (13.1% survival)
- Disparity: 7.4x difference