Titanic EDA – Comprehensive Findings

Dataset: train.csv (n = 891)

Reference: EDA.ipynb (see notebook for plots)

Summary of Key Findings

- Overall survival: 342/891 (38.4%).
- Sex and Pclass are the strongest drivers of survival.
- Fare and Cabin-known are strong supporting signals.
- Embarked has a weaker but noticeable effect (C > Q ≈ S).
- Age differences are modest overall; distribution is non-normal with missingness.
- Small families survive more than solo; large families the worst.

Missing Values

Counts and percentages of null entries by column (percent over total rows).

Column	Missing	Percentage
Cabin	687	77.1
Age	177	19.87
Embarked	2	0.22

- Age missing: 177 (19.87%).
- Embarked missing: 2 (0.22%).
- Cabin missing: 687 (77.10%).

Outlier Scan – IsolationForest

Unsupervised outlier detection on numeric features (excluding target). Reported are counts and ranges for key fields among flagged outliers.

- Flagged outliers: 214 rows (24.0%).
- Fare among outliers – min: 7.25, median: 36.88, max: 512.33.
- Age among outliers – min: 0.75, median: 33.5, max: 80.0.

Numeric Relationships & Distribution Shape

Spearman correlations for all numeric pairs (rank-based; values near ±1 indicate stronger monotonic association).

Var1	Var2	SpearmanR
Pclass	Fare	-0.688
SibSp	Parch	0.45
SibSp	Fare	0.447
Parch	Fare	0.41
Pclass	Age	-0.362
Survived	Pclass	-0.34
Survived	Fare	0.324
Age	Parch	-0.254
Age	SibSp	-0.182

Survived	Parch	0.138
Age	Fare	0.135
Survived	SibSp	0.089
PassengerId	SibSp	-0.061
Survived	Age	-0.053
Pclass	SibSp	-0.043
PassengerId	Age	0.041
PassengerId	Pclass	-0.034
Pclass	Parch	-0.023
PassengerId	Fare	-0.014
PassengerId	Survived	-0.005
PassengerId	Parch	0.001

Numeric summary: central tendency, spread, and shape (skew, kurtosis) for each numeric variable.

Variable	Mean	Median	Std	Skew	Kurtosis
PassengerId	446.0	446.0	257.354	0.0	-1.2
Survived	0.384	0.0	0.487	0.479	-1.775
Pclass	2.309	3.0	0.836	-0.631	-1.28
Age	29.699	28.0	14.526	0.389	0.178
SibSp	0.523	0.0	1.103	3.695	17.88
Parch	0.382	0.0	0.806	2.749	9.778
Fare	32.204	14.454	49.693	4.787	33.398

Categorical Associations vs Survived (Cramer's V)

Association strength (0 to 1) between each categorical variable and Survived; higher = stronger association.

Variable	CramersV
Sex	0.54
Pclass_lbl	0.337
Cabin_known	0.313
Ticket	0.311
Embarked	0.166

Pclass

Distribution of passenger classes (First/Second/Third) with percentages.

Pclass	Count	Percentage
Third	491	55.11
First	216	24.24
Second	184	20.65

Survival rate within each class (survivors/total and percentage).

Pclass_lbl	Survivors	Total	SurvivalRate%

First	136	216	63.0
Second	87	184	47.3
Third	119	491	24.2

Survival by Pclass × Sex (class gradient persists within sex).

Pclass_lbl	Sex	Survivors	Total	SurvivalRate_percent
First	female	91	94	96.8
First	male	45	122	36.9
Second	female	70	76	92.1
Second	male	17	108	15.7
Third	female	72	144	50.0
Third	male	47	347	13.5

Sex

Sex distribution and survival rates (female survival markedly higher).

Sex	Count	Percentage
male	577	64.76
female	314	35.24

Sex	Survivors	Total	SurvivalRate%
female	233	314	74.2
male	109	577	18.9

Age

Distributional statistics, missingness, and normality tests for Age.

Metric	Value
Missing	177.0
Missing%	19.87
Available	714.0
Mean	29.7
Median	28.0
Std	14.53
P25	20.12
P75	38.0

- Shapiro-Wilk: statistic=0.981, p-value=7.337e-08 (alpha=0.05).
- Anderson-Darling: statistic=3.823, critical@5%=0.783.

Age by Survived (group means/medians and sample sizes).

Survived_lbl	MeanAge	MedianAge	N
No	30.63	28.0	424
Yes	28.34	28.0	290

- ANOVA: F=4.271, p=0.039 (mean difference).

- Kruskal-Wallis: H=1.970, p=0.160 (median difference).

family_size (engineered)

Engineered from SibSp+Parch: Solo(0), Small(1–2), Large(≥3).

family_size	Count	Percentage
Solo	537	60.27
Small Family	263	29.52
Large Family	91	10.21

Survival by family_size (rates highlight small-family advantage).

family_size	Survivors	Total	SurvivalRate%
Large Family	31	91	34.1
Small Family	148	263	56.3
Solo	163	537	30.4

Fare

Distributional statistics and normality tests for Fare (heavy right skew).

Metric	Value
Missing	0.0
Mean	32.2
Median	14.45
Std	49.69
Min	0.0
P90	77.96
P95	112.08
Max	512.33
Skew	4.79
Kurtosis	33.4

- Shapiro-Wilk: statistic=0.522, p-value=1.084e-43.
- Anderson-Darling: statistic=122.170, critical@5%=0.784.

Fare by Survived (mean/median differences reflect class effects).

Survived_lbl	MeanFare	MedianFare	N
No	22.12	10.5	549
Yes	48.4	26.0	342

Mean Fare by Sex × Pclass × Survived (survivors pay more within strata).

Sex	Pclass_lbl	No	Yes
female	First	110.6	105.98
female	Second	18.25	22.29
female	Third	19.77	12.46
male	First	62.89	74.64

male	Second	19.49	21.1
male	Third	12.2	15.58

Cabin (known vs unknown)

Binarized cabin availability and corresponding survival rates.

Cabin_known	Count	Percentage
unknown	687	77.1
known	204	22.9

Cabin_known	Survivors	Total	SurvivalRate%
known	136	204	66.7
unknown	206	687	30.0

Embarked

Port distribution (S, C, Q) including NAs and survival by port.

Embarked	Count	Percentage
S	644	72.28
C	168	18.86
Q	77	8.64
nan	2	0.22

Embarked	Survivors	Total	SurvivalRate%
C	93	168	55.4
Q	30	77	39.0
S	217	644	33.7

Survival by Sex × Embarked (female advantage consistent across ports).

Sex	Embarked	Survivors	Total	SurvivalRate_percent
female	C	64	73	87.7
female	Q	27	36	75.0
female	S	140	203	69.0
male	C	29	95	30.5
male	Q	3	41	7.3
male	S	77	441	17.5