Titanic Survival Analysis - Full Results and Interpretation

Overview

This project uses SAS tools to predict survival on the Titanic using demographic and travel data. Models include logistic regression and random forest. We analyze survival by class, sex, family, and location.

Data Summary

Train dataset size: 714 observations
 Survival values:

- 0 = Did not survive

- 1 = Survived

Pclass, Embarked, and Survival

proc freq table: Pclass by Embarked controlling for survival

Pclass refers to ticket class:

- Pclass=1 is first class

- Pclass=2 is second class

- Pclass=3 is third class

Table: Pclass Survival Breakdown

Pclass	Survived	Died	Total	Death Rate (%)
1	120	64	184	34.8
2	83	90	173	52.0
3	85	270	355	76.1

Explanation:
First-class passengers survived more because they were on upper decks, had early access to lifeboats, and were prioritized during evacuation. Third-class passengers stayed deep below deck and faced crowded escape routes.

Survival by Family Size

proc freq table: FamilySize by Survived

FamilySize = SibSp + Parch + 1

- SibSp = siblings/spouses aboard

- Parch = parents/children aboard

Table: Family Size vs Survival

Family Size	Survived	Died	Total
1	130	274	404
2	76	63	139
3	53	40	93
4	21	6	27
5	3	8	11
6	3	19	22
7	4	8	12
8	0	6	6

Interpretation:small
Passengers with small families (2-4 members) had better survival rates not just because of group support, but also because many were in Pclass 1 or 2, located on upper decks. This gave them faster access to lifeboats and crew assistance. The combination of manageable family size and higher class location made escape more likely compared to large families in lower decks who struggled to evacuate together.

Survival by IsAlone

IsAlone = 1 if FamilySize = 1, else 0

Table: IsAlone vs Survival

IsAlone	Survived	Died	Total
0	160	150	310
1	130	274	404

Interpretation:
Passengers who traveled alone (IsAlone = 1) had the highest death rate, with 274 out of 404, or 68 percent, not surviving. Most solo travelers, about 56 percent, were in Pclass 3, located on the lowest deck with poor lifeboat access. Being alone likely made it harder to act quickly or receive help during the emergency. This suggests that being with others improved communication, coordination, and access to safety, which increased the chance of survival.

Logistic Regression Results

proc logistic using: Survived = sex age pclass embarked familysize isalone

Observations used: 712
 AIC (model quality): 650.464
 C-statistic (AUC): 0.859 (very good)

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.9	Somers' D	0.718
Percent Discordant	14.0	Gamma	0.719
Percent Tied	0.1	Tau-a	0.346
Pairs	122112	c	0.859

Significant Predictors:

Variable	p-value	Odds Ratio	Effect
Sex	<.0001	13.67	Females had much higher chance
Pclass = 1	<.0001	10.72	First class more likely to survive
Pclass = 2	<.0001	3.20	Second class better than third
Age	<.0001	0.959	Older passengers less likely to survive
FamilySize	0.0031	0.727	Larger family reduced survival slightly

Not significant:

- Embarked (port boarded)

- IsAlone

Odds Ratio Estimates		
Effect	Point Estimate	95% Wald Confidence Limits
Sex female vs male	13.673	8.849 21.126
Age	0.959	0.944 0.975
Pclass 1 vs 3	10.720	5.974 19.235
Pclass 2 vs 3	3.200	1.959 5.226
Embarked C vs S	1.470	0.861 2.512
Embarked Q vs S	0.701	0.237 2.074
FamilySize	0.727	0.589 0.898
IsAlone	0.612	0.338 1.109

Random Forest Results

proc hpforest on clean_train datasetF

Used 100 trees
 Best performance between 70-100 trees
 Average square error (OOB) = 0.145

Variable Importance (based on MSE reduction)

Variable	Importance Score
Sex	0.056

Pclass   0.023

FamilySize   0.007

IsAlone   0.004

Embarked   0.004

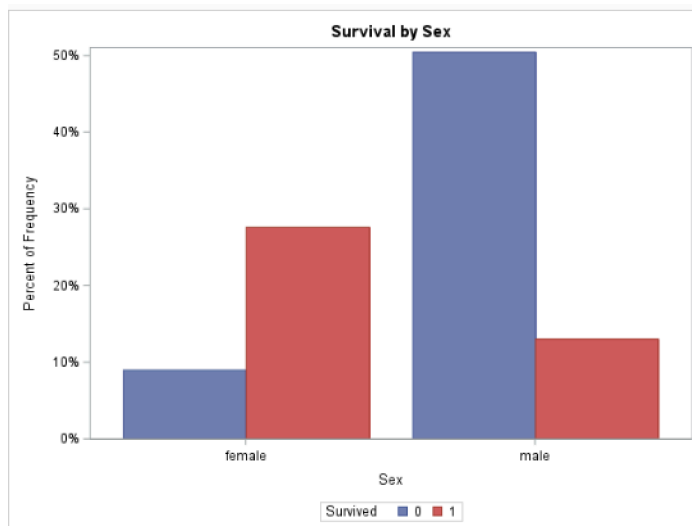Age   0.001

Interpretation:
 Sex and Pclass were the most important factors. Family relationships also helped improve accuracy.

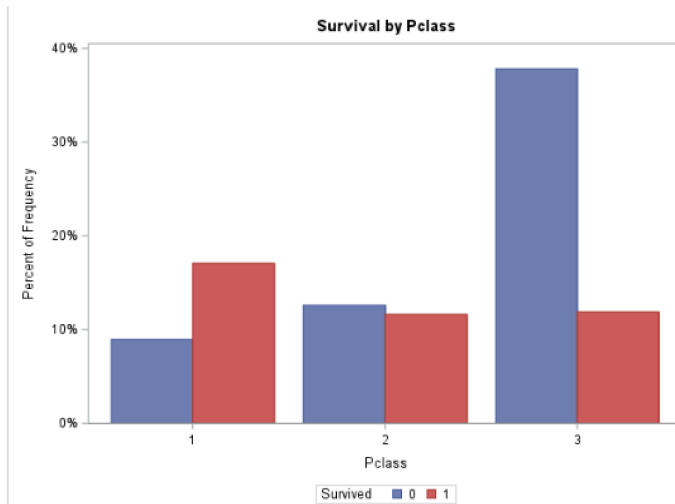# Visuals: proc sgplot

Bar Chart: Sex vs Survival

proc sgplot created grouped bar chart
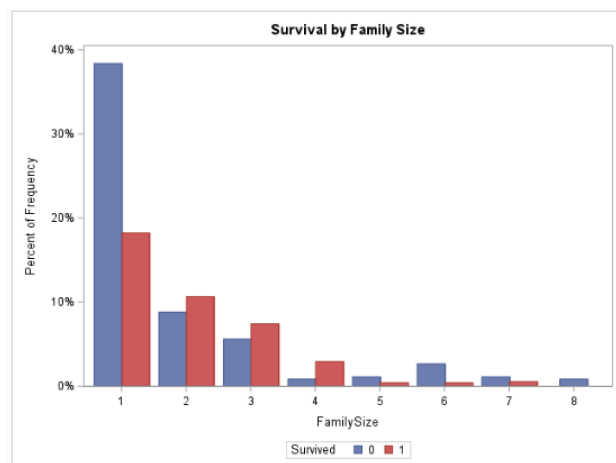 Result: Females had much higher survival

Bar Chart: Pclass vs Survival

proc sgplot clustered bar chart shows lower survival for third class

Survival by Pclass

Bar Chart: Family Size vs Survival

Survival by Family Size

Bar Chart: IsAlone vs Survival

Final Notes

The analysis clearly shows that class, gender, and family connections strongly impacted survival on the Titanic. These patterns reflect both the physical structure of the ship and social priorities during evacuation.

- Pclass was a major survival divider. First-class passengers had the highest survival rate (65.2%), while third-class passengers had the lowest (23.9%). This suggests a need to closely monitor and improve emergency access and escape plans for lower-class or economy passengers on modern vessels. Crew members and protocols should ensure equal evacuation access across all deck levels.

- Sex was the most powerful predictor. Women had 13.67 times higher odds of survival than men. This reflects the "women and children first" rule being enforced during lifeboat loading. From a modern perspective, it shows how priority-based rescue orders and crew decisions can drastically affect survival outcomes.

- Family structure mattered. Solo travelers had worse outcomes, while medium-sized families (2-4) had better odds. Travel groups may help each other navigate and escape faster. Cruise safety plans might consider family zones or pairing systems to avoid isolation.

- AUC (Area Under Curve) = 0.859 in the logistic regression model shows strong predictive power. This high AUC score indicates the model can reliably distinguish between survivors and non-survivors, making it valuable for real-world applications like:

    - Evacuation training simulations

    - Pre-screening risk profiles for safety drills

    - Predictive modeling in airline, rail, or ferry transport

- The Random Forest model confirmed the same key drivers: sex, Pclass, and family structure. Sex and class remained most important. The model's consistent accuracy across 70–100 trees adds confidence to the results.

Recommendation

In future disaster planning, special attention should be given to economy-class and solo travelers. Systems should be tested to ensure no group is unintentionally delayed due to location, crowding, or access. Strong predictive models like these allow companies or governments to simulate and prepare for evacuation under different scenarios.