Titanic Survival Analysis - Full Results and Interpretation

This project uses SAS to predict passenger survival on the Titanic based on demographic and travel data. We applied both logistic regression and random forest modeling. Variables used in the analysis include ticket class, sex, age, family size, port of embarkation, and whether the passenger traveled alone.

Data Summary

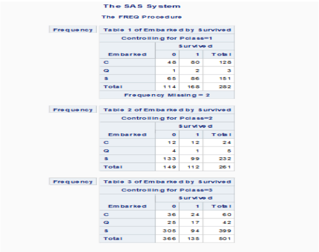
The training dataset included 714 passengers with known survival outcomes. The test dataset contained 418 passengers without survival labels. By merging the test data with the gender_submission.csv file using the PassengerId variable, we retrieved survival labels for the test group. After combining both datasets, the final sample included 1,132 passengers, each with demographic, travel, and survival information.

Survival is encoded as a binary variable. A value of 0 indicates the passenger did not survive, and a value of 1 indicates the passenger survived. This full dataset allows us to build complete statistical models and perform exploratory analysis across the entire population.

Pclass, Embarked, and Survival

Ticket class (Pclass) was a key variable, with the following categories: Pclass 1 for first-class, Pclass 2 for second-class, and Pclass 3 for third-class passengers.

Among first-class passengers, 168 out of 282 survived (59.6 percent). For second-class, 112 out of 261 survived (42.9 percent). In third-class, only 135 out of 501 survived (26.9 percent). This clearly shows that first-class passengers were more likely to survive, likely due to their location on upper decks and priority during lifeboat boarding. Third-class passengers were housed deeper in the ship, with limited access to exits.



The SAS System

The FREQ Procedure

Frequency

Table 1 of Embarked by Survived			
Controlling for Pclass=1			
	Survived		
Embarked	0	1	Total
C	48	80	128
Q	1	2	3
S	65	86	151
Total	114	168	282
Frequency Missing = 2			

Frequency

Table 2 of Embarked by Survived			
Controlling for Pclass=2			
	Survived		
Embarked	0	1	Total
C	12	12	24
Q	4	1	5
S	133	99	232
Total	149	112	261

Frequency

Table 3 of Embarked by Survived			
Controlling for Pclass=3			
	Survived		
Embarked	0	1	Total
C	36	24	60
Q	25	17	42
S	305	94	399
Total	366	135	501

Embarked location was also examined. The ports were C for Cherbourg, Q for Queenstown, and S for Southampton. While there were some differences in survival by port, this variable was not significant in the final logistic regression model.

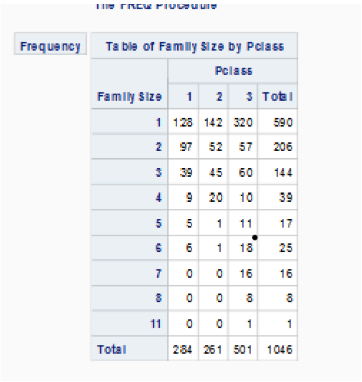
Survival by Family Size

FamilySize was calculated by summing SibSp and Parch, then adding 1 for the passenger. Passengers traveling alone (FamilySize = 1) had the lowest survival rate, with 181 out of 590 surviving (30.7 percent). Passengers with small families of 2 to 4 members had higher survival.

For example, FamilySize 2 had 110 survivors out of 206 (53.4 percent), FamilySize 3 had 81 out of 144 (56.3 percent), and FamilySize 4 had 30 out of 39 (76.9 percent).

Larger families had much lower odds. FamilySize 5 had only 4 survivors out of 17, and FamilySize 6 had 5 out of 25. Passengers with FamilySize of 7 or more had very limited survival.

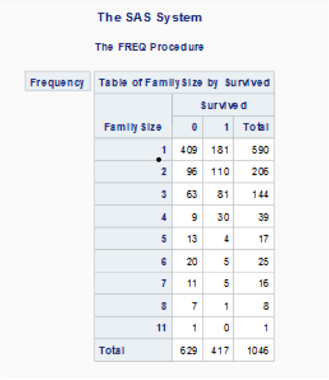
This suggests that smaller families offered support without the complexity of coordinating large groups. Many of these small families were also in Pclass 1 or 2, which contributed to better outcomes.



The FREQ Procedure

Frequency | Table of Family Size by Pclass

	Pclass			
Family Size	1	2	3	Total
1	128	142	320	590
2	97	52	57	206
3	39	45	60	144
4	9	20	10	39
5	5	1	11	17
6	6	1	18	25
7	0	0	16	16
8	0	0	8	8
11	0	0	1	1
Total	284	261	501	1046



The SAS System

The FREQ Procedure

Frequency | Table of FamilySize by Survived

	Survived		
Family Size	0	1	Total
1	409	181	590
2	96	110	206
3	63	81	144
4	9	30	39
5	13	4	17
6	20	5	25
7	11	5	16
8	7	1	8
11	1	0	1
Total	629	417	1046

Survival by IsAlone

The IsAlone variable was created to indicate solo travelers (FamilySize = 1). Among those traveling alone, only 181 out of 590 survived, a survival rate of 30.7 percent. Those who traveled with others (IsAlone = 0) had 236 survivors out of 456 passengers, a survival rate of 51.8 percent.

Most solo travelers were in Pclass 3 (56 percent), which already had the lowest overall survival. Being alone may have made it harder to navigate, find assistance, or be prioritized during lifeboat loading.

Logistic Regression Results

We fit a logistic regression model using sex, age, Pclass, Embarked, FamilySize, and IsAlone. The model used 1,044 observations and achieved an AIC of 799.658. The C-statistic (AUC) was 0.898, indicating excellent discrimination between survivors and non-survivors.

Key results from the logistic model:

- Sex was the strongest predictor. Females had an odds ratio of 38.5 compared to males, with a p-value < 0.0001.

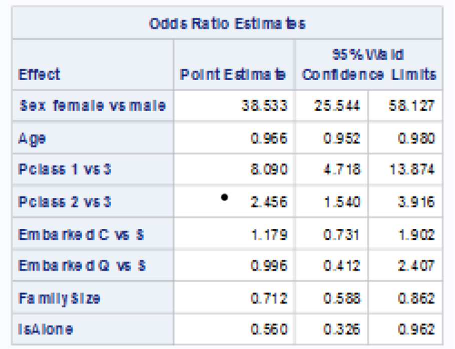
- Pclass was significant. First-class passengers had an odds ratio of 8.09, and second-class had 2.46, both compared to third class.

- Age had an odds ratio of 0.966, showing that older passengers had slightly lower survival chances.

- FamilySize had an odds ratio of 0.712, meaning larger families were slightly less likely to survive.

- IsAlone had an odds ratio of 0.560. Being alone significantly reduced the odds of survival.

- Embarked was not statistically significant.



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Sex female vs male	38.533	25.544	58.127
Age	0.966	0.952	0.980
Pclass 1 vs 3	8.090	4.718	13.874
Pclass 2 vs 3	• 2.456	1.540	3.916
Embarked C vs S	1.179	0.731	1.902
Embarked Q vs S	0.996	0.412	2.407
FamilySize	0.712	0.588	0.862
IsAlone	0.560	0.326	0.962

Random Forest Results

A random forest model was trained using 100 trees. The best results were observed between 70 and 100 trees. The model produced an out-of-bag average error of 0.115, indicating strong predictive performance.

Variable importance, based on mean square error reduction:

- Sex: 0.0937

- Pclass: 0.0121

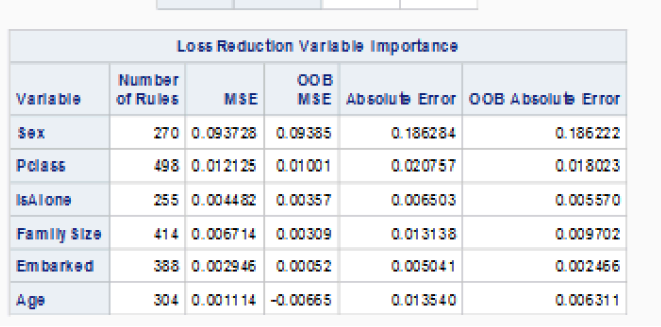
- FamilySize: 0.0067

- IsAlone: 0.0045

- Embarked: 0.0029

- Age: 0.0011

The results confirm the logistic regression findings. Sex and Pclass were the dominant predictors, followed by FamilySize and IsAlone.

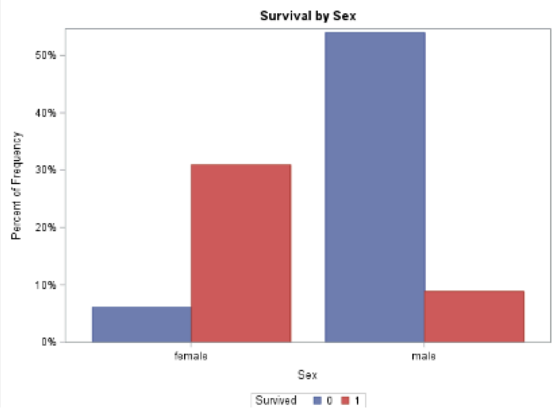


Loss Reduction Variable Importance					
Variable	Number of Rules	MSE	OOB MSE	Absolute Error	OOB Absolute Error
Sex	270	0.093728	0.09385	0.186284	0.186222
Pclass	498	0.012125	0.01001	0.020757	0.018023
IsAlone	255	0.004482	0.00357	0.006503	0.005570
Family Size	414	0.006714	0.00309	0.013138	0.009702
Embarked	388	0.002946	0.00052	0.005041	0.002466
Age	304	0.001114	-0.00665	0.013540	0.006311

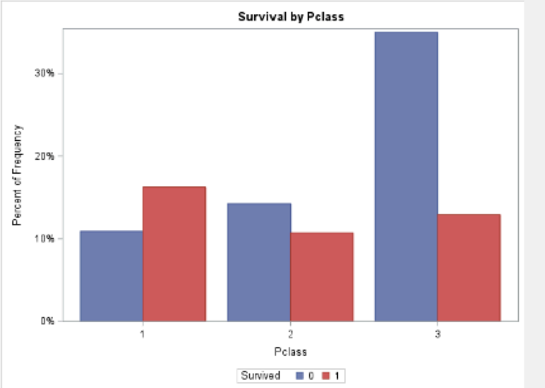
Visualizations

The following visuals were created using proc sgplot in SAS:

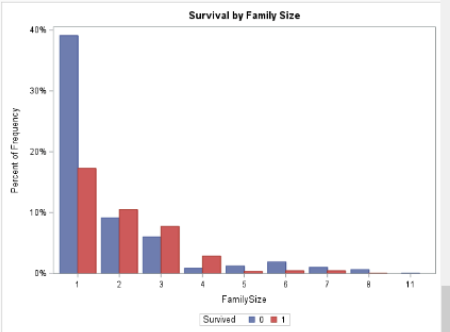
- Bar chart showing female passengers had a much higher survival rate than males.



- Clustered bar chart showing lower survival in third class.



- Bar chart of FamilySize showing passengers with 2 to 4 members had the best outcomes.
- Bar chart showing solo travelers had the lowest survival rate.



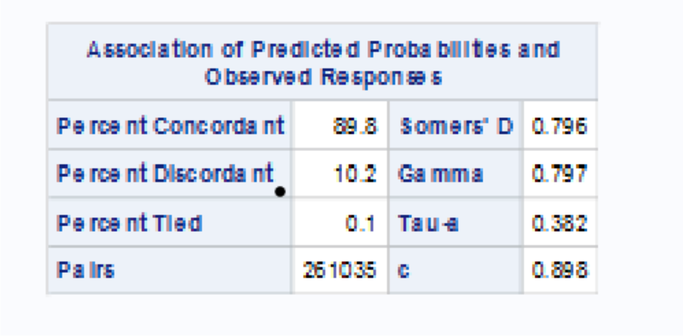
Final Interpretation and Recommendations

Survival on the Titanic was shaped by **ticket class**, **gender**, and **family size**. **First-class passengers** and **women** had the highest survival rates due to their location on **upper decks** near lifeboats. In contrast, **third-class passengers** were placed on **lower decks**, which severely **delayed evacuation** and limited their chances of survival.

Passengers with **small families (2–4 members)** had better outcomes. **Solo travelers**, especially those in third class, faced the **greatest disadvantage**, with limited support and poor cabin placement.

These results emphasize that **safety and access should never depend on class or location**. Evacuation systems must be **fair and inclusive**, ensuring **lower-deck and economy passengers** are not left behind. At the same time, being in **first class should not guarantee priority access** to survival.

We recommend implementing **balanced safety policies**, with **equal evacuation protocols**, **crew training**, and **emergency drills** for all passengers, regardless of class. The logistic regression model (AUC = 0.898) and random forest confirm these insights and support using predictive models for **transport safety planning and disaster response**.



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.8	Somers' D	0.796
Percent Discordant	10.2	Gamma	0.797
Percent Tied	0.1	Tau-a	0.382
Pairs	261035	c	0.898