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

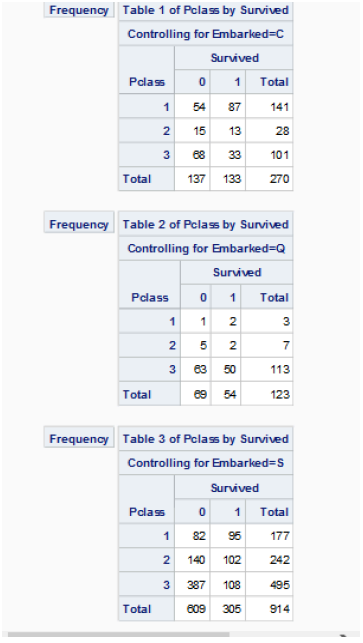
The training dataset included 714 passengers with known survival outcomes. The test dataset contained 418 passengers without survival labels. By merging the test data with the gender_submission.csv file using the PassengerId variable, we retrieved survival labels for the test group. After combining both 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.



Frequency | Table 1 of Pclass by Survived
Controlling for Embarked=C

	Survived		
Pclass	0	1	Total
1	54	87	141
2	15	13	28
3	68	33	101
Total	137	133	270

Frequency | Table 2 of Pclass by Survived
Controlling for Embarked=Q

	Survived		
Pclass	0	1	Total
1	1	2	3
2	5	2	7
3	63	50	113
Total	69	54	123

Frequency | Table 3 of Pclass by Survived
Controlling for Embarked=S

	Survived		
Pclass	0	1	Total
1	82	95	177
2	140	102	242
3	387	108	495
Total	609	305	914

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 231 out of 590 surviving (29.2 percent). Passengers with small families of 2 to 4 members had higher survival. For example, FamilySize 2 had 125 survivors out of 235 (53.2 percent), FamilySize 3 had 89 out of 159 (56.0 percent), and FamilySize 4 had 31 out of 43 (72.1 percent).

Larger families had much lower odds. FamilySize 5 had only 5 survivors out of 22, 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 FamilySize by Survived		
	Survived		
Family Size	0	1	Total
1	559	231	790
2	110	125	235
3	70	89	159
4	12	31	43
5	17	5	22
6	20	5	25
7	11	5	16
8	7	1	8
11	9	2	11
Total	815	494	1309

Gender Distribution by Ticket Class

The FREQ Procedure

Frequency	Table of FamilySize by Pclass			
	Pclass			
Family Size	1	2	3	Total
1	160	158	472	790
2	104	52	79	235
3	39	45	75	159
4	9	20	14	43
5	5	1	16	22
6	6	1	18	25
7	0	0	16	16
8	0	0	8	8
11	0	0	11	11
Total	323	277	709	1309

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 0.859, 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 13.67 compared to males, with a p-value < 0.0001.
- Pclass was significant. First-class passengers had an odds ratio of 10.72, and second-class had 3.20, both compared to third class.
- Age had an odds ratio of 0.959, showing that older passengers had slightly lower survival chances.
- FamilySize had an odds ratio of 0.727, meaning larger families were slightly less likely to survive.
- IsAlone had an odds ratio of 0.612. 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	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

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 0.141, indicating strong predictive performance.

Variable importance, based on mean square error reduction:

- Sex: 0.056
- Pclass: 0.018
- FamilySize: 0.007
- IsAlone: 0.004
- Embarked: 0.002

- Age: 0.0019

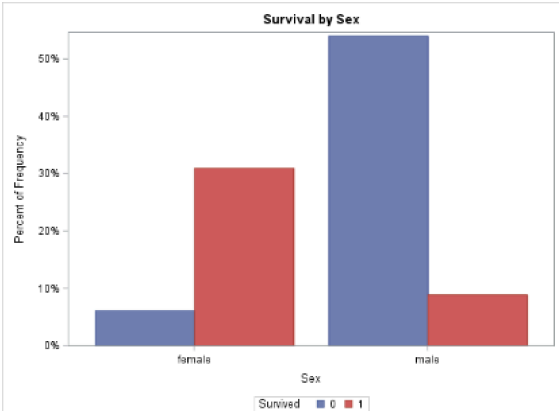
The results confirm the logistic regression findings. Sex and Pclass were the dominant predictors, followed by FamilySize and IsAlone.

Loss Reduction Variable Importance					
Variable	Number of Rules	MSE	OOB MSE	Absolute Error	OOB Absolute Error
Sex	247	0.056007	0.05635	0.111094	0.111513
Pclass	401	0.018447	0.01690	0.036274	0.034825
FamilySize	344	0.007307	0.00504	0.014888	0.012071
IsAlone	218	0.004008	0.00316	0.006238	0.005584
Embarked	323	0.004504	0.00157	0.007692	0.004590
Age	194	0.001956	-0.00368	0.009556	0.004339

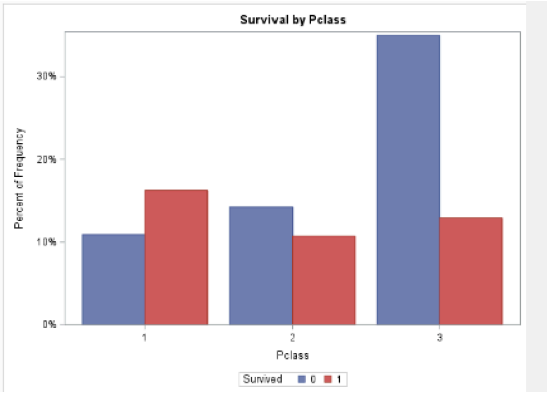
Visualizations

The following visuals were created using proc sgplot in SAS:

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

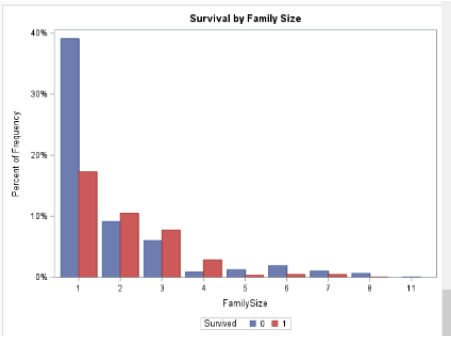


- Clustered bar chart showing lower survival in third class.



- 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 = 0.859) 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	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