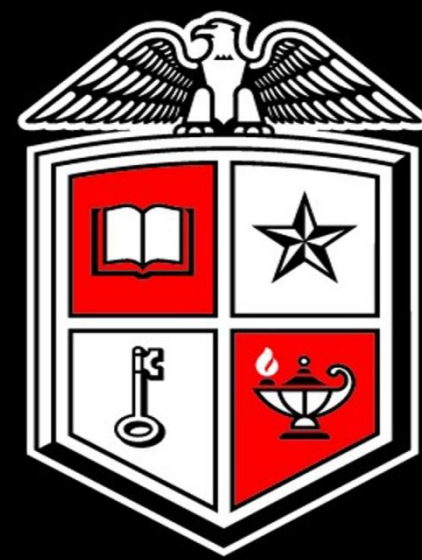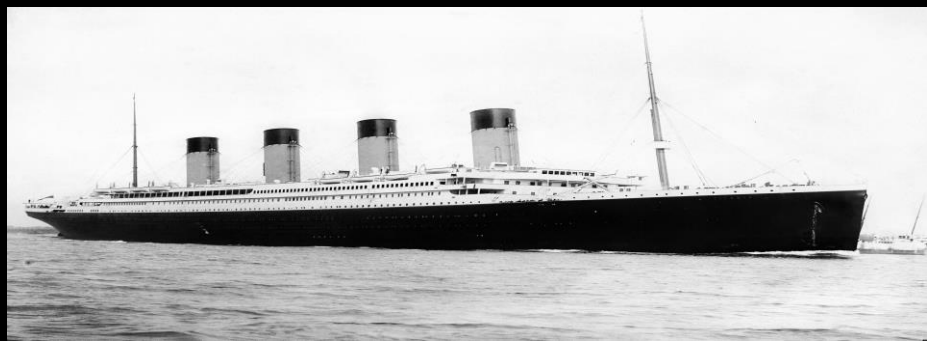# Predicting the Survival of Titanic Passengers

By: Ashley Fowler
MATH 5365 Project, Dr. Volchenkov
Texas Tech Department of Mathematics

## Brief Overview

The sinking of the RMS Titanic occurred on the night of 14 April through to the morning of 15 April 1912 in the North Atlantic Ocean, four days into the ship's maiden voyage from Southampton to New York City. The largest passenger liner in service at the time, Titanic had an estimated 2,224 people on board when she struck an iceberg at around 23:40 (ship's time) on Sunday, 14 April 1912. Her sinking two hours and forty minutes later at 02:20 (05:18 GMT) on Monday, 15 April resulted in the deaths of more than 1,500 people, which made it one of the deadliest peacetime maritime disasters in history. This tragic event led to the creation of numerous safety regulations and policies to prevent such a catastrophe from happening again. Some critics, however, argue that circumstances other than luck resulted in a disproportionate number of deaths.

## Methodology

Titanic data from Kaggle.com
- 10 variables
- R Programming

### Data Dictionary

| Variable | Definition | Key |
|---|---|---|
| Survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| Sex | Gender | |
| Age | Age in years | |
| sibsp | # of siblings / spouses aboard the Titanic | |
| parch | # of parents / children aboard the Titanic | |
| Ticket | Ticket number | |
| fare | Passenger fare | |
| cabin | Cabin number | |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

## Research Objective

The purpose of this analysis is to explore factors that influenced a person's likelihood to survive.

### Research Question

What covariates are significant factors in association to the survival of passengers on the Titanic?

## Statistical Analysis

**Exploratory data analysis** - An approach to analyzing data sets to summarize their main characteristics, often with visual methods
- Relationship Between Dependent and Independent Variables
- Relationship to Survival Rate

### Decision Trees

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
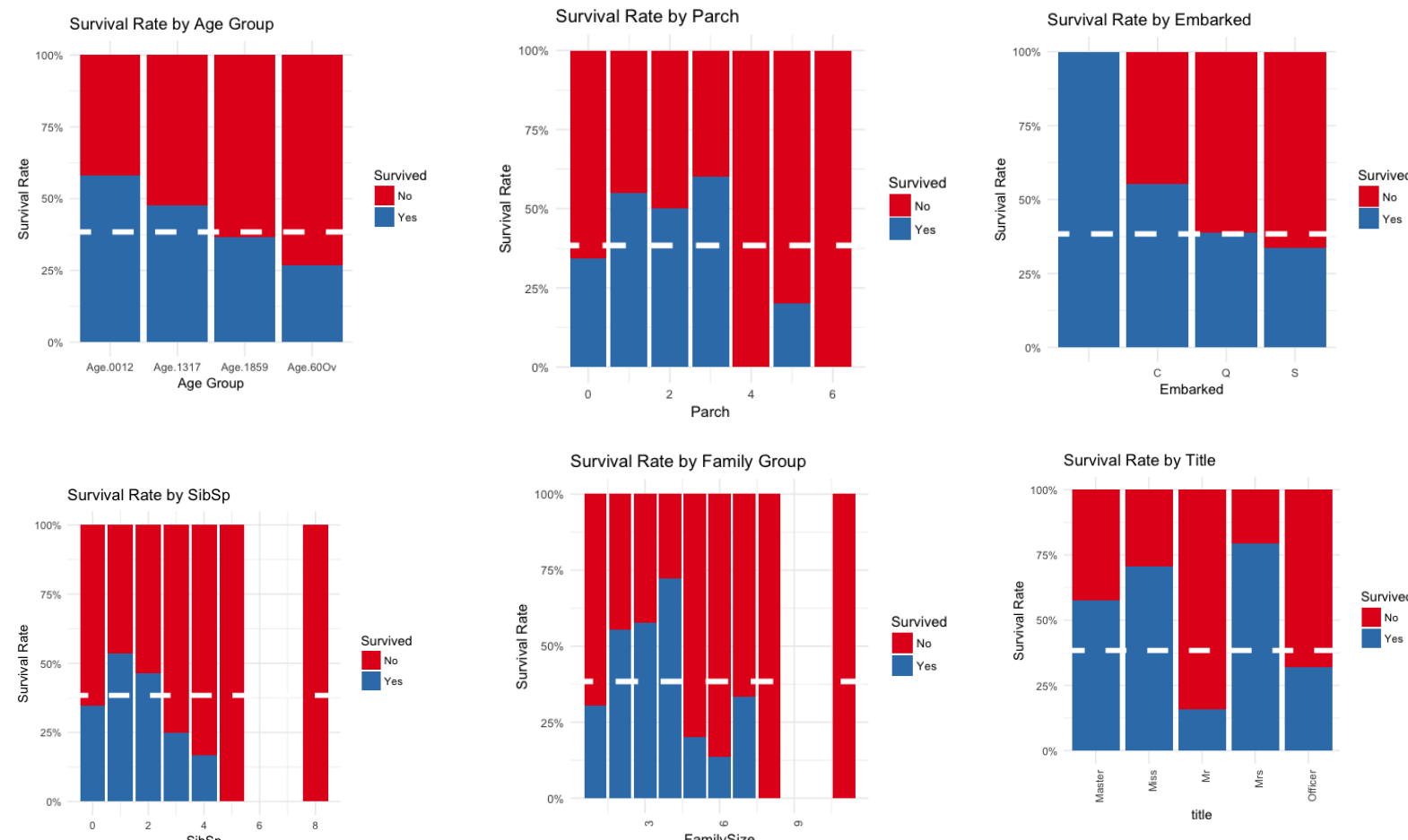
## Results

### Descriptive Tables

The independent variable, Survived, is labeled as a Bernoulli trial where a passenger or crew member surviving is encoded with the value of 1. Among observations in the train set, approximately 38% of passengers and crew survived.

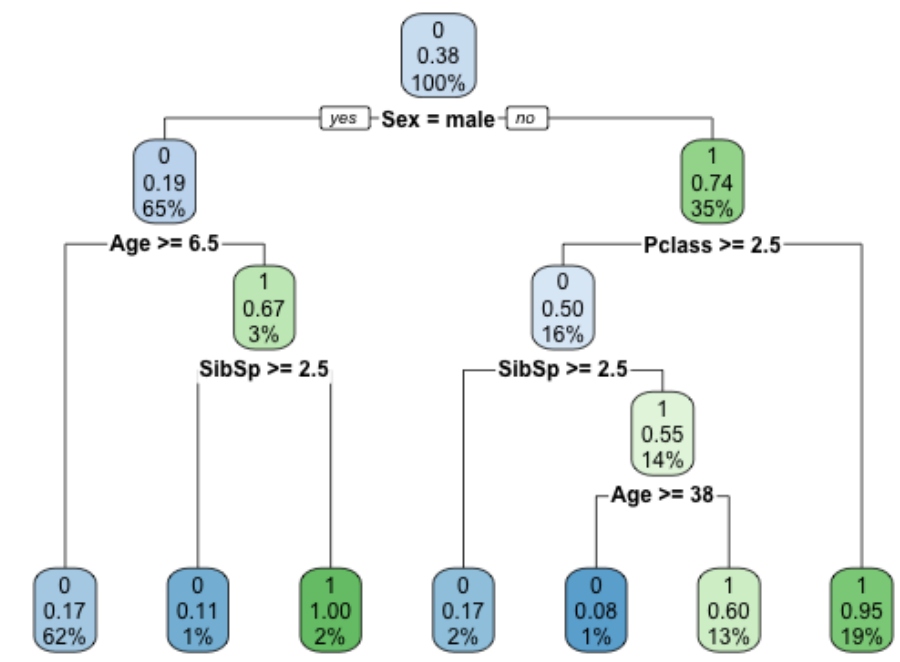| Survived | n | freq |
|---|---|---|
| No | 549 | 0.6161616 |
| Yes | 342 | 0.3838384 |

### Gender Characteristics

| Female | Male |
|---|---|
| 314 | 577 |

## Results — Two-way comparison on the number of males and females that survived

| | Didn't Survive | Survived |
|---|---|---|
| Female | 26% | 74% |
| Male | 81% | 19% |

## Relationship to Survival Rate



1 = 1st, 2 = 2nd, 3 = 3rd Class

## Relationship to Survival Rate



## Decision Tree



For this decision tree, the subset looks at the Pclass, gender, age and number of siblings. For example, you can see that if you were a male older than 6.5 years old, you had a 67% change to survive.

## Discussion/Conclusion

This project aimed to predict the survival of titanic Passengers based off certain factors. It seems as though women and children had the best chance to survive. It was also discovered that there was a lower Survival rate for Females of class 3.
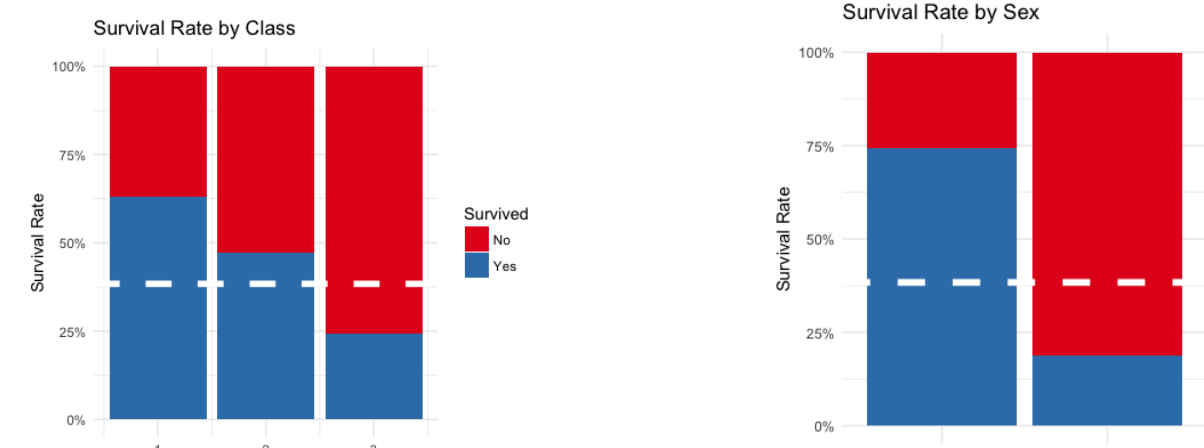
### Future Research

There is still a huge amount of analysis to be done. Future research includes creating more decision trees and also doing a Random Forest analysis. More variables can also be made from the data that can also be analyzed.