**Final Project**

The goal of this report is to provide comprehensive analysis that will help predict the survival rate of passengers by gender. Passengers ID, P class, Name, Sex, Age, Parch, Ticket fare, Cabin and embarked were used to analyze the effect on finding their survival rates. Logistic fit, Residuals and mosaic plots were used to determine the results.

**Process:**

Data Gathering: Data provided was collected from online which included a range of factors that can possibly influence the gender and survival of passengers.

Data cleaning: Choose the best model that uses only significant factors Our aim is to gain a better understanding of the factors that influenced survival, including passenger demographics, ticket class, and other variables.

A screenshot of a computer

Description automatically generated with medium confidence

Result analysis: We will begin by exploring the dataset, which contains information on passengers who were aboard the ship. This will involve examining the distribution of variables, identifying missing data, and cleaning the data where necessary. Once we have a clean dataset, we will perform exploratory data analysis to gain insights into the relationships between different variables. Next, we will build a predictive model to determine which factors were most important in predicting survival. We will use logistic regression and mosaic plots, a common technique for binary classification problems, to model the probability of survival based on passenger characteristics. Finally, we will evaluate the performance of our model and draw conclusions about the factors that influenced survival on the ship.

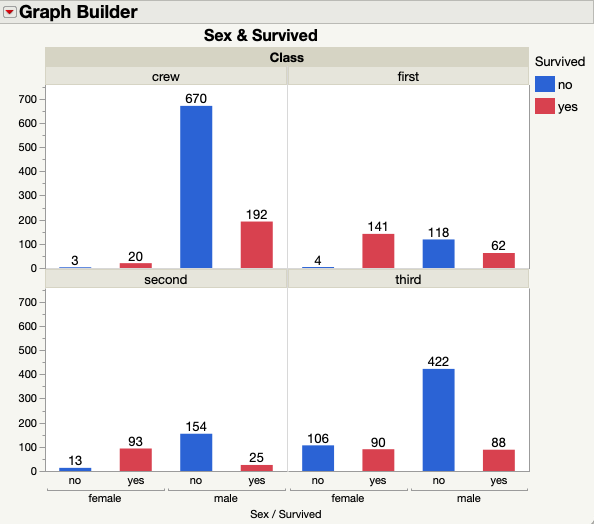
Let’s first look at what features we have. There are some features that we do not want to use. For example, Passenger Id, because that is simply assigned as the order of which each passenger data is entered into this dataset There might be something useful in the Name column, for example, maybe we can tell which passengers are part of a family, or if you’ve noticed, every passenger seems to have a title “Mr.”, “Miss”, etc. the Sex column can be easily used as a binary feature because there are only 2 unique values, so we can use 0 and 1 to represent male and female. Age we can now use after we’ve filled in the missing values, same with Fare. Sib Sp, parch are also numeric features we can use it.

Therefore, having ready-to-use data, let's consider a potential relationship we could graph. We will only use "Survived," "P class," "Sex," and "Age" in today's analysis. We are undoubtedly attempting to forecast certain aspects of survivability. We could infer correlations between several variables intuitively. For instance, we could assume that older persons and newborns have a lower chance of surviving due to their physical weakness or immobility. Let’s perform analysis on “Sex” and “Survived” first. Since both variables are nominal, we should use Categorical Data Analysis to perform modeling.

To display how frequently each Sex-Survived pair occurs. There can be four possible pairs: Males survived, males didn't, females did, and females didn't. Choose "Analyze" from the menu bar on JMP to build such a table. Choose "Fit Y by X" from the dropdown menu.

Graphical user interface

Description automatically generated with medium confidence



## Text Description automatically generated

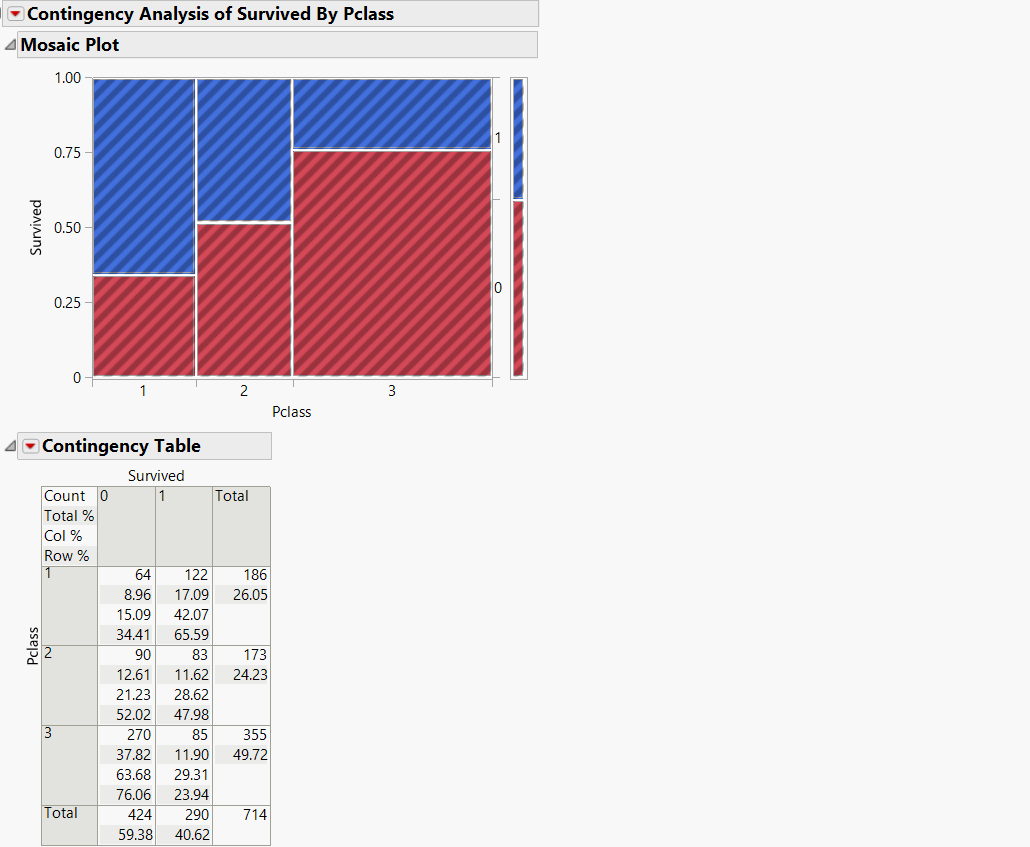
The graph compares the percentage of survivors against the percentage of fatalities based on "Sex," in detail. Based on our sample data, it is obvious that more women than men survived the crash, with a substantially higher percentage of women surviving than men. The Contingency table, which is immediately below the graph, contains the precise numbers and percentages. The table can be easily understood. The four integers that correlate to row "female" and column "0" indicate: 64 females, or 8.96% of the sample size, did not survive, making up the total. A total of 15.09% of the individuals in our sample did not survive, and this percentage is made up entirely of females.

H0= Two factors are independent

Ha = Two factors are not independent

The better Prob>Chi Sq, the smaller it is. If the Prob>Chi Sq number is less than 0.05, you may be 95% certain that the claim "female survival rate is greater than male survival rate" is accurate and 5% certain that it is untrue. If you're familiar with p-values, you'll know that.0001 is a very low p-value, indicating that there is very little chance that the chi square value will exceed the empirical value of your data (assuming the null hypothesis is true, H0). In other words, the likelihood that males have a higher survival rate than females do not exist very often, less than 0.0001. As a result, we draw the conclusion that "Sex" has a substantial role in predicting survival.

Currently, we can run a similar test to compare "P class" to "Survived," which would result in the output seen below.



Let's compare "Age" and "Survived" to conclude. As "Age" is continuous and "Survived" is nominal, we should compute the model using logistic regression.

Modeling the chance of success or failure using logistic regression. In this instance, success is maintained. We are interested in learning whether "Age" affects survivability in any way or whether people who are older or younger are more likely to survive. Go to Analyze- Like what we did earlier. Set "Age" to be X and "Survived" to be Y, then fit Y by X. If you click "OK," the plot should appear below.

Scatter chart

Description automatically generated

As we can see, the logistic function is not very significant intercept has very high p-value. The points are scattered all over the x-y plane. Thus, we can’t reject the null hypothesis. However, failure to reject null doesn’t mean that “Age” is not related to “Survived”. In other words, the likelihood that males have a higher survival rate than females do not exist very often, less than 0.0001. As a result, we draw the conclusion that "Sex" has a substantial role in predicting survival.