Dennis Houlihan

Assignment 7: Titanic Simulation

The sinking of the Titanic in 1912 remains one of history's most infamous tragedies. However, the detailed records of its passengers offer a unique opportunity for data analysis. A previous report examined a dataset of 891 individuals aboard the ship to determine what were the key factors that influenced survival rates. More specifically, the role of age, biological sex, ticket price, ticket class, and cabin designation on survival rates were explored. Taking these into account, we attempted to create a model to simulate the sinking of the ship and predict the death count for the whole ship.

In order to predict the amount of deaths, we had to synthesize a full dataset to represent all of the passengers and crew aboard the Titanic. A full record of how many people were on the ship is not available, so we assume a normal distribution of people with a mean of 2,224 and a standard deviation of 400. The majority of the people aboard were passengers, with an estimated ratio of 0.59 of passengers to crew. This ratio is based upon the commonly estimated number of 885 crew members and 1317 passengers (SOURCE). For each iteration of the simulation, a random number of people were drawn given the distribution above, and the ratio of 0.59 was used to determine the number of passengers and crew specifically.

The dataset that was previously analyzed contained only passengers and would not be suitable for predicting deaths among crew members. In order to do so, we took a dataset from the Titanic Encyclopedia (SOURCE) which had a dataset of crew members aboard. For the crew, we based the simulation on age and gender as they did not have tickets nor a specific cabin or class assigned. We expect to have a better estimation of passenger deaths as there are more variables being accounted for in their simulations, including age, gender, ticket price, ticket class, and cabin designation.

The simulation starts by constructing a logistic regression model that trains and validates on the passenger and crew datasets independently. It takes a random subset comprising 30% of the total dataset to be the training data while the other 70% is used for validation. It then provides a survival probability to each person based upon the factors inputted to the model. To initially assess the model's effectiveness, we compared the average predicted survival probabilities to the actual outcomes across each demographic group and key factor analyzed.

We first compared the model on the passenger dataset, which is outlined in Table 1. When comparing the rates by ticket price, we see that overall, it matched quite well with a weighted mean percent error of -1.23% and a weighted mean absolute (ABS) percent error of 8.86%. There was quite a lot of variance between different age groups, but that could be due to lower counts yielding more unreliable results for those particular groups. The model closely matched the rate when we compared by ticket class, with a weighted mean percent error of only -4.20% and a weighted mean ABS percentage error of 4.98%. When comparing the rates by gender, we see close agreement with percent differences of -0.68% and -2.73% for female and

male survivor rates, respectively. Finally, the model performed well when comparing with age ranges with a mean percentage error of -0.21% and a mean ABS percentage error of 7.71%.

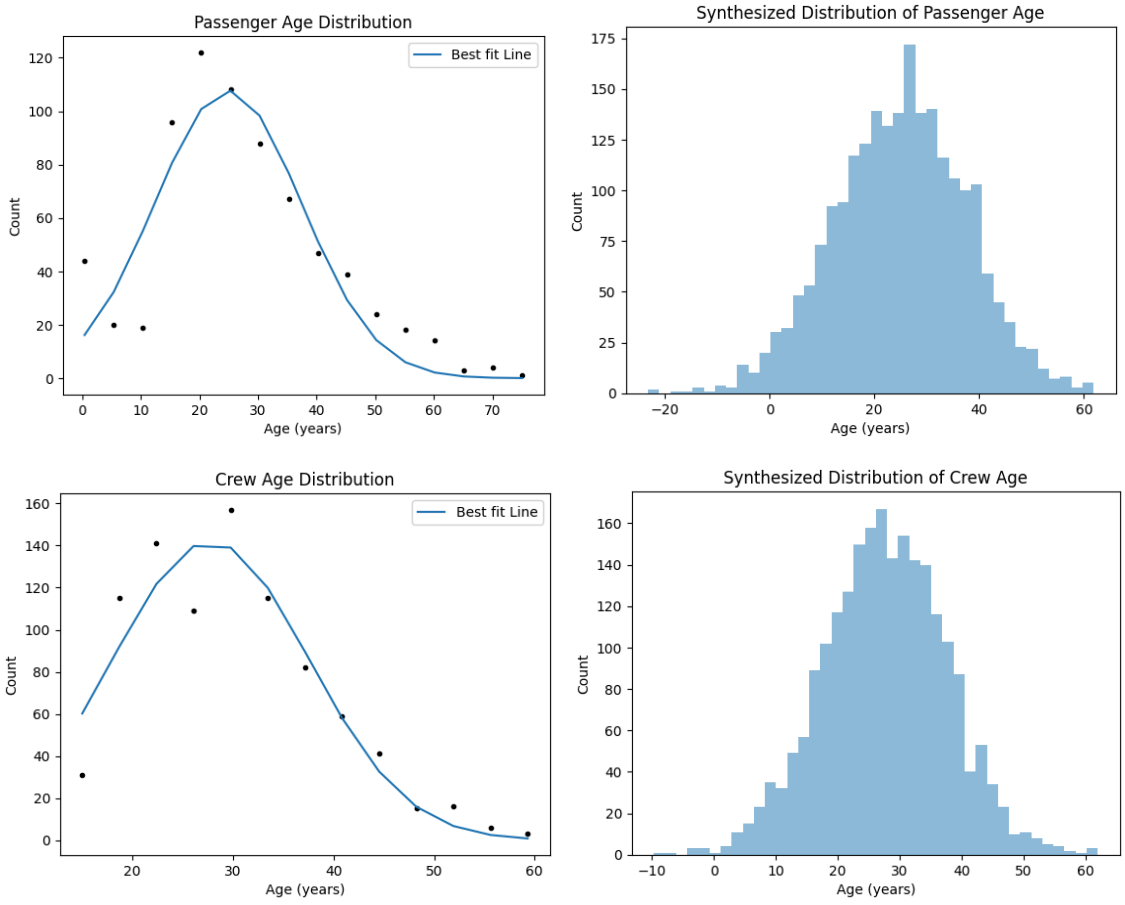
The model also performed well with the crew dataset. Table 2 outlines the comparisons for crew survival rates. There was a mean percentage error of only -2.73% and -0.68% for female and male survival rates, respectively. When comparing with age ranges, there was a weighted mean percentage error of -5.28% and a weighted mean ABS percentage error of 12.09%. Almost all of the percentage errors for both passenger and crew fell within 10% and all were within 15%. Given these results, we felt comfortable using the model to predict survival rates beyond the datasets given.

Categories	Actual Passenger (%)	Predicted Passenger (%)	Weighted Mean Percent Error (%)	Weighted Mean Absolute Percent Error (%)
Male	18.89	19.41	-2.73	-
Female	74.20	74.71	-0.68	-
Age	-	-	-5.28	12.09
Ticket Price	-	-	-1.23	8.86
Ticket Class	-	-	-4.20	4.98

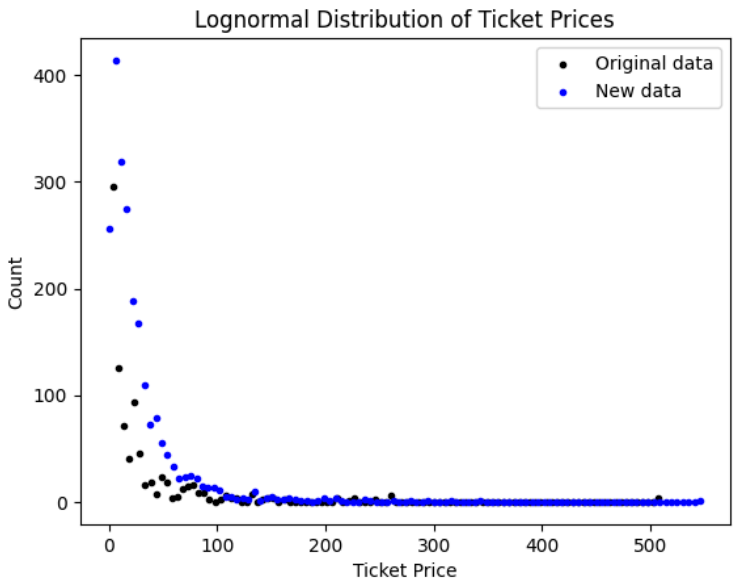
Categories	Actual Crew (%)	Predicted Crew (%)	Weighted Mean Percent Error (%)	Weighted Mean Absolute Percent Error (%)
Male	22.03	22.88	-2.73	NA
Female	86.96	79.31	-0.68	NA
Age	-	-	-5.28	12.09

Since the model provides a survival rate based on a logistic regression approach, which correlates the rate of survival with factors given such as age and ticket price, constructing a synthetic dataset that follows the distributions seen in the actual datasets is very important. Again, the total number of passengers and crew is determined by a random sample of a normal distribution with a mean of 2,224 and a standard deviation of 400, further split into passenger and crew designations with a static ratio provided above.

To provide an age for each synthetic person on board, we first fit a function to the age distribution of the passenger and crew datasets. For both, a simple Gaussian, or normal, distribution was fit to a histogram of ages with a bin width of 5 years. This approach is not perfect, as it does not account well for the spike of children within the passenger dataset (ages < 5 years old). Figure 1 shows the age distributions of passengers (top left) and crew (bottom left) with the best fit line as well as a synthesized distribution of passenger (top right) and crew (bottom right) age with N = 2224.



To synthesize genders, we simply assigned genders based upon the male-to-female ratio in the respective datasets. The same procedure was done for ticket classes. To assign ticket prices to the synthetic passengers, we fit a lognormal distribution to the ticket prices given in the passenger dataset. Figure 2 shows the original distribution of ticket prices from the dataset as well as the synthesized distribution with N = 2224 and with a bin width of $5.

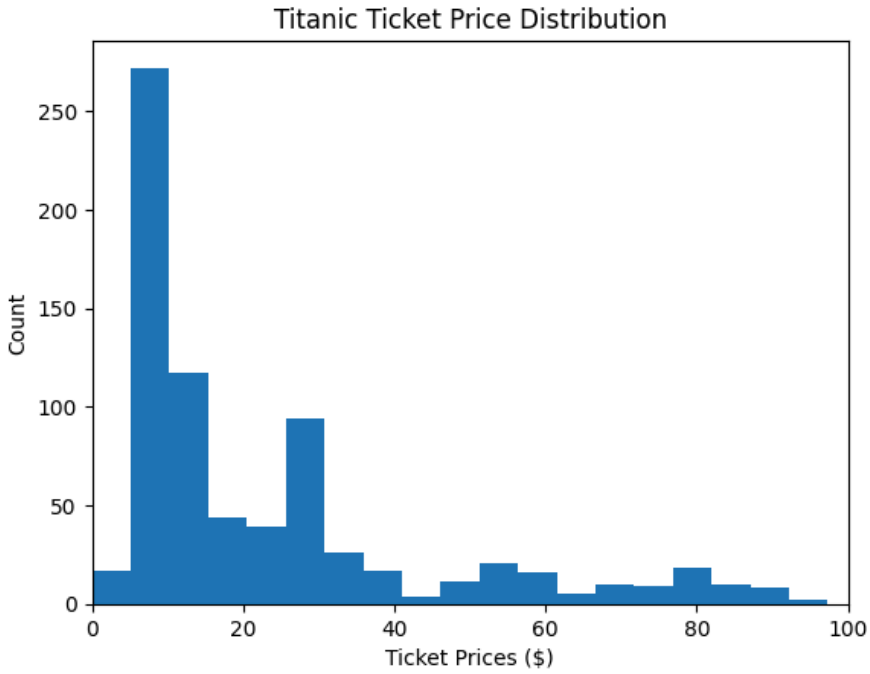


The Titanic dataset contains 891 passengers with varying demographics, class distinctions, and survival outcomes. The average ticket price for passengers in this dataset was $32.20 with a median ticket price of $14.45. Figure 1 shows the distribution of ticket prices for passengers in the dataset. Most tickets bought were below $20 but there are smaller bumps around $30, $55, and $80. These could represent different tiers of tickets corresponding to cabin types or positions.

Once the synthetic dataset is constructed, it is time to run the simulation. We used the Python package ‘simpy’ to perform this simulation. For each person, their survival probability was compared with a random number between 0 and 1. If the random number was less than the survival probability, then the person survived, and if it was greater, then that person died. We ran the simulation ten times, with an average of 2299 total people, 1509 deaths, and 790 survivors. Our 95% confidence interval for deaths was between 1284 and 1735 people, while the survivors interval was between 677 and 902. This yielded an average survival rate of 34.4% with a 95% confidence interval of 34.1% to 34.7%. While the total number of survivors and nonsurvivors is unknown, it is widely accepted that around 1,500 people died while around 706 people survived. This yields a percent difference of only 0.30% for nonsurvivors and 5.60% for survivors. This does show that the model may overestimate the number of survivors, however the percent error is still well within 10%.

To continue this assignment, we redid our analysis from homework 5 and the simulation, but this time with the ‘titanic-modified.csv’ file. Below are our results.

The Titanic dataset contains 789 passengers with varying demographics, class distinctions, and survival outcomes. The average ticket price for passengers in this dataset was $33.06 with a median ticket price of $14.50. Figure 3 shows the distribution of ticket prices for passengers in the dataset. Most tickets bought were below $20 but there are smaller bumps around $30, $55, and $80. These could represent different tiers of tickets corresponding to cabin types or positions.



There are 6 types of cabins, labelled with prefixes ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, and ‘T’. The survival rates were 43%, 76%, 62%, 79%, 74%, and 0%, respectively. There were also 187 survivors and 413 non-survivors (~31% survival rate) without a cabin designation, so this needs to be accounted for when doing the analysis. Figure 4 shows how many survivors and non-survivors there were for each cabin along with the survival rates. As you can see, despite most passengers not surviving, most of the passengers with a cabin designation did survive. Given this fact, it is difficult to extrapolate the numbers to the total dataset. However, we can still compare survival rates between the cabins. Cabins with prefixes B, D, and E have a significantly higher survival rate than the others, with cabin A and C having much lower survival rates. Cabin T only has one total passenger, so it is difficult to extrapolate any conclusions from it.

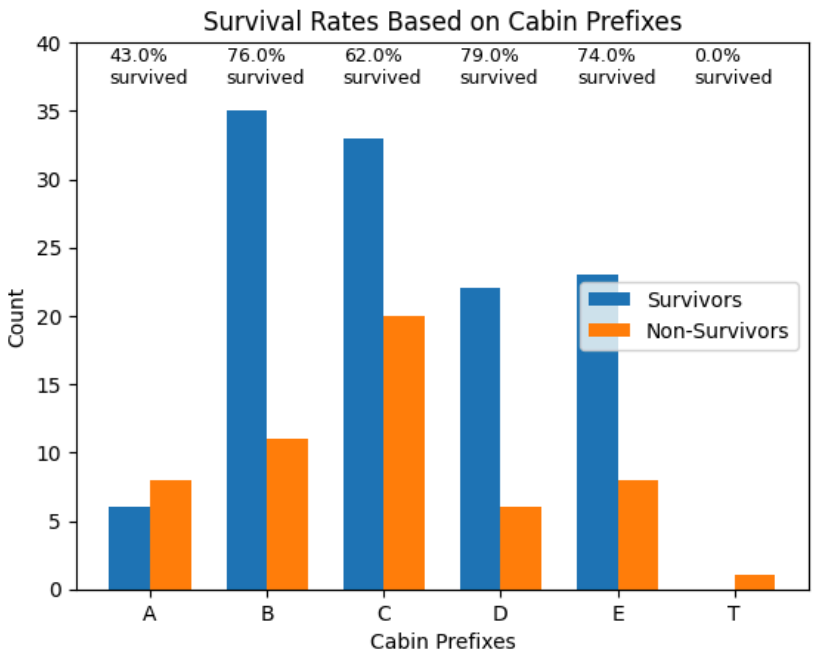
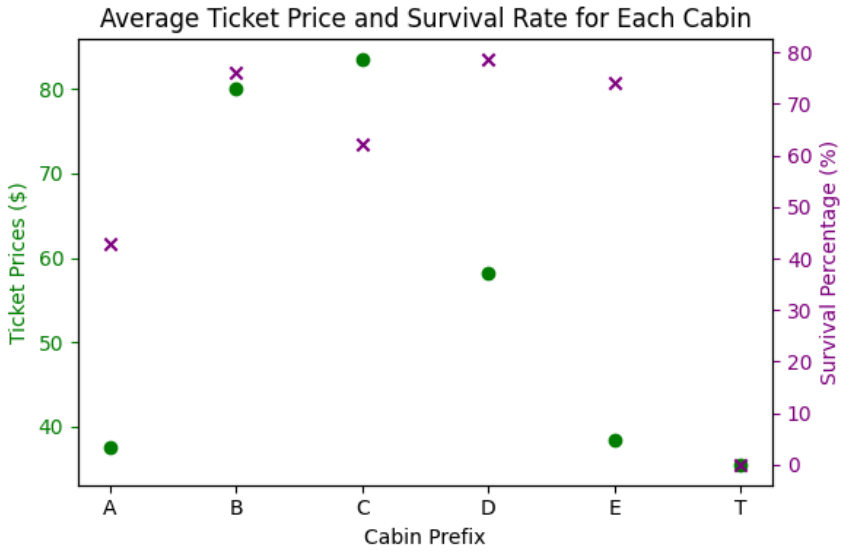
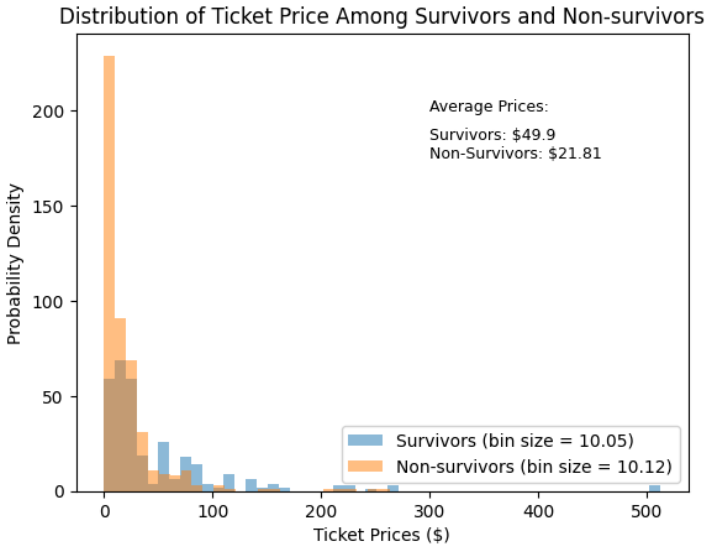


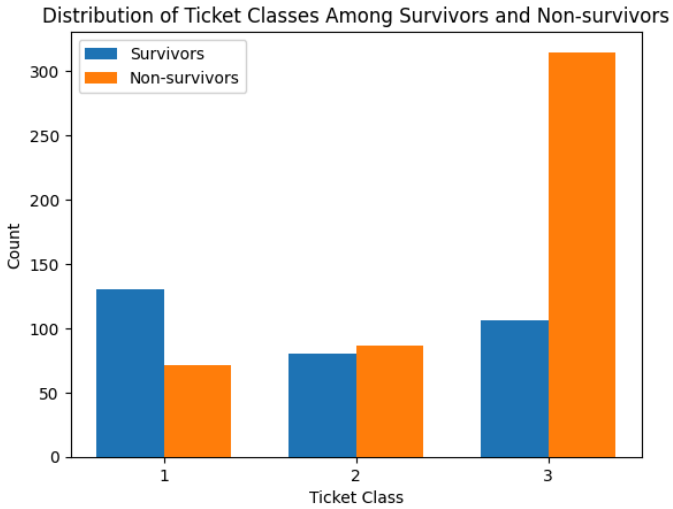
Figure 5 goes further to see if there was any correlation between a cabin's average ticket price and the survival rate. Cabins A, D, and E cost less on average while B and C have the highest average ticket prices. There was no clear link between the average ticket prices and survival rate. Cabins D and E had the highest survival rate while also costing the least and cabin B cost the most on average but has a lower survival rate than B, D, and E. However, given that the average ticket price for someone without a cabin designation was $19.54 and the survival rate was 31%, it is possible that cabins themselves were cost-prohibitive and most people without a cabin designation were in more vulnerable positions on the ship.



This correlation between survival rate and ticket prices is further supported by Figure 6. It shows ticket price distributions for both survivors and non-survivors. Most non-survivors had lower-priced tickets, while there were very few non-survivors with tickets above $100. We can also see the average ticket price for a survivor was $49.90 and $21.81 for non-survivors. This could be explained by the fact that more expensive tickets afforded passengers better cabins, access to lifeboats, or access to other resources to escape the flooding water.



Furthermore, we can look at how ticket class impacted the survival rate of the passengers. There are three distinct classes labelled ‘1’, 2’, and ‘3’ as seen in Figure 7. Class 3 showed the biggest discrepancy, as there was only about a 25% survival rate. The other classes had a more even split between survivors and non-survivors. This could indicate that passengers with class 3 tickets were housed in more compromising places within the ship, or perhaps did not have as much access to lifeboats or other resources.



Exploring the demographics of the dataset could also reveal more factors that influenced survival rate. Of the passengers, 64.13% were male with an average age of 31.05 years while 35.87% were female with an average age of 28.46 years. The average age of survivors was 28.59 years and 31.2 years for non-survivors, which does not show any correlation with age and survival rate. Figure 8 may explain why as it shows the distribution of age for all passengers on the ship. When one looks at the distribution, it makes sense then that the average ages of survivors and non-survivors would not yield a good description since most passengers were between the ages of 20 and 40. But if we dig a bit deeper and plot the distributions of age for survivors and non-survivors, we begin to see a pattern.

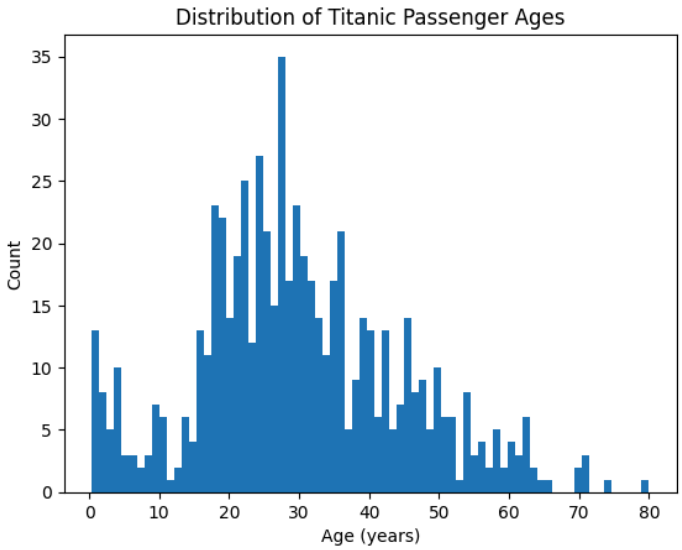
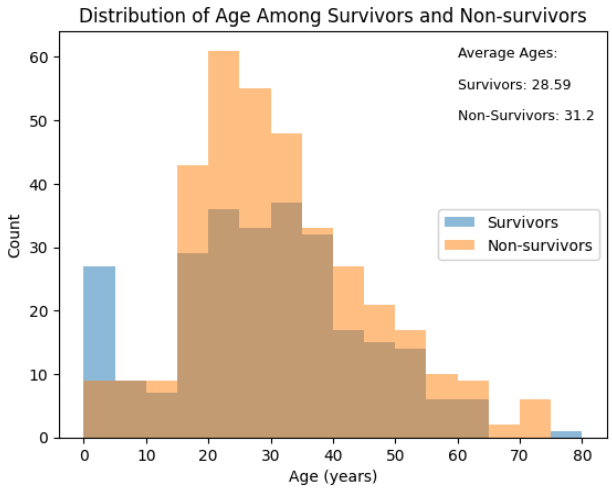
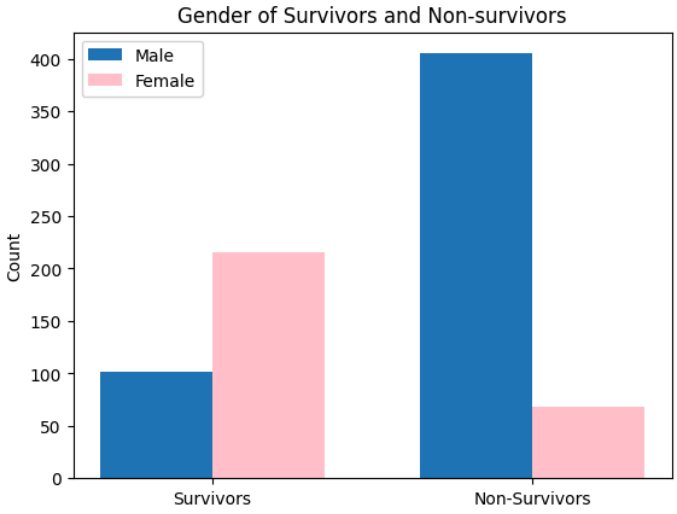


Figure 9 shows those exact distributions. As you can see, most survivors and non-survivors were between 20 and 40, but we can see some other noticeable features. The older demographics tended to have lower survival rates. This could be because they were less able to maneuver and compete for lifeboat spots, or if they were in the water, their bodies were less likely to be able to withstand the cold. We also see a spike in survival rate with children aged 5 and below. This could be that during evacuations, children were prioritized for lifeboat spots.



Along with age, the biological sex of the passengers had a big impact on survival rate. Figure 10 shows a histogram that outlines the gender breakup for both survivors and non-survivors. Most survivors were women while most non-survivors were men. This could be because during evacuations, both women and children were prioritized.



We again ran the simulation ten times, this time with the modified dataset. There was an average of 2199 total people, 1430 deaths, and 769 survivors. Our 95% confidence interval for deaths was between 1271 and 1587 people, while the survivors interval was between 680 and 858. This yielded an average survival rate of 34.97% with a 95% confidence interval of 34.7% to 35.2%. While the total number of survivors and nonsurvivors is unknown, it is widely accepted that around 1,500 people died while around 706 people survived. This yields a percent difference of only 8.9% for survivors and -4.60% for nonsurvivors. This does show that the model may overestimate the number of survivors and underestimate the number of

nonsurvivors, however the percent error is still well within 10%. Overall, it did not perform as well as the original dataset we used.