Exploratory Analysis of Titanic Data

DATASCI W200, Spring 2022 | Tues 6:30 PM PST
Team Members: Daphne Lin, Yujin Kim

Github Repository

https://github.com/UC-Berkeley-I-School/Project2_Lin_Kim

Introduction

The Titanic was a legendary ship its owners claimed was unsinkable. Though the sinking of the Titanic was an unfortunate disaster, we can learn much about the characteristics of those most likely to have survived and those most likely to have perished through a passenger log pieced together by researchers.

Data Sources

Our primary dataset consists of the survival status of individual passengers on the Titanic and passenger characteristics. This dataset was merged with the table-formatted data identified from two Titanic Lifeboat data sources that captured the boat number, boat deployment time, boat location, and boat capacity. Further details are captured in the tables below.

Source 1: Titanic Data	
Source	http://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pdf
Abstract	Data consists of the survival status for individuals who boarded the Titanic. Per site, "[the] titanic data frame does not contain information from the crew, but it does contain actual ages of half of the passengers. The principal source for data about Titanic passengers is the Encyclopedia Titanica."
Size	117 KB; 1309 rows x 14 columns

Source 1: Columns of Interest	
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survived	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex of Individual
age	Age of Individual

Sources 2 and 3: Titanic Lifeboats Launch Time, Boat Launch Side, and Capacity	
Sources	https://titanicfacts.net/titanic-lifeboats/ https://www.titanic-titanic.com/titanics-lifeboats/
Abstract	Data was captured in a table for each of the lifeboats launched, organized by order of launch time beginning from 12:40 AM to 02:15 AM, and whether the lifeboat was on the starboard or port side and capacity. We joined Source 2 and 3 on Source 1 using the boat column as the key.

Sources 2 and 3: Columns of Interest	
boat	Lifeboat Number/Letter
boat_deploy_time	Each lifeboat's deployment time.
boat_loc	Boat Location (Starboard; Port)
capacity	Each boat's max capacity

Data Cleaning

See Appendix A for null and non-null counts.

Upon reviewing the data definitions and null values, we decided to fill in or clean the following:

- **Family:** Combined sibsp and parch into one column named 'family'
- **Child_women_men**: Using the columns age and sex, we added a column named 'child_women_men'. We defined a child as under 18 female and under 13 male based on historic accounts. We defined women as 18 and above female, and men as 13 and above male.
- **Fare:** Filled in the 1 null fare value with the mean of that individual's pclass (pclass 3 with an average fare of 13.30)
- **Age**: Filled in missing ages with the mean age of each individual's pclass. Rounded up <1 year old infants to 1 year of age to account for pd.cut bin handling of 0.
- **Boat**: After reviewing the unique boat field entries, we identified individuals with multiple boat numbers (e.g., '5 7' or 'C D'). After reviewing the data, entries with '5 7', 'C D', and '13 15' were entries of married couples. We made the assumption that the couple was split between two lifeboats and updated the data, randomly choosing which passenger received which boat number (e.g., for a married couple with entry '5, 7' we allocated Mrs -> 5; Mr -> 7). There were two entries with '13 15' and one entry with '13 15 B' in the boat field. We made the assumption that one of these individuals was on each of the boats 13, 15, and B, and allocated this randomly. For the rest of the multiple boat numbers '5 9', '8 10' and '15 16' there was only one entry each, so we randomly selected a boat number for each of these individuals from their listed boat numbers.
- **Gender:** For purposes of creating a correlation heat map later on, we replaced all males with 0 and all females with 1.
- **Cabin**: We dropped the cabin column due to 77% missing values.
- **Body**: We dropped the body column as those who did not survive may not have had their body recovered. We decided to focus on the survival column as the indicator of survival.
- **Ticket**: We dropped the ticket column after reviewing the unique values which were not consistent in formatting and pattern. We decided to focus on the fares and the pclass columns as the indicator of socioeconomic status.
- **Home.dest**: We dropped the home.dest column due to 43% missing values.

Basic Characteristics of Titanic Data

See Appendix B for histograms of passenger age, women/men/children distribution, pclass, embarked, and morbidity.

- 1309 passengers total
- 394 women, 122 children, and 793 men
- Ages ranging from 1 to 80
- Pclass 1: 323 passengers, Pclass 2: 277 passengers, Pclass 3: 709 passengers
- Embarked from: S: 914 passengers, C: 207 passengers, Q: 123 passengers
- **Morbidity: Of 1309 passengers, 500 survived, and 809 died.** Breakout of survivor counts by children, women and men are shown below:

		survived
child_woman_man	**survived**	
child	1	77
	0	45
man	0	659
	1	134
woman	1	289
	0	105

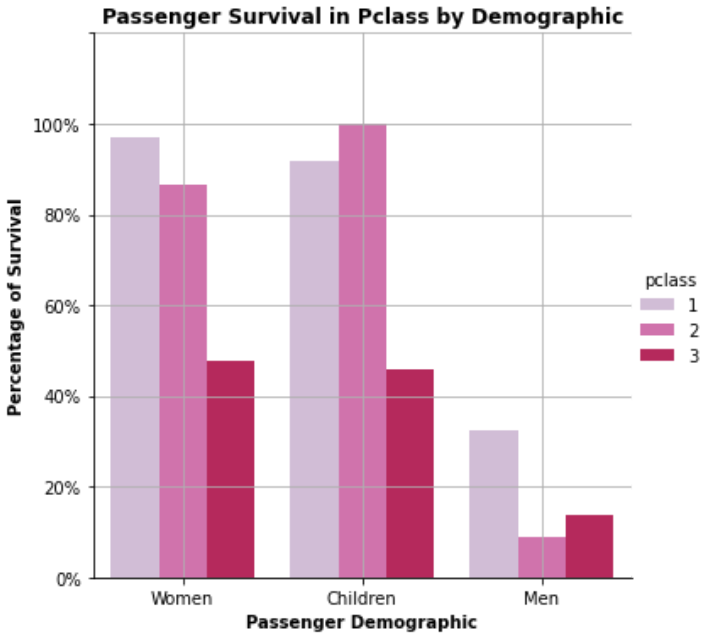
- **Number of Children, Women, and Men on Lifeboat at Each Lifeboat Deployal Time vs. Number of Children, Women, and Men Remaining on Titanic after that round of Boat Deployal**

Boat Deployal Times	Number of Children on Lifeboat	Number of Men on Lifeboat	Number of Women on Lifeboat	Number of Children Remaining on Titanic Post Boat Deployal	Number of Women Remaining on Titanic Post Boat Deployal	Number of Men Remaining on Titanic Post Boat Deployal
12:43 AM	0	12	11	122	383	781
12:45 AM	1	12	14	121	369	769
01:00 AM	4	12	33	127	336	757
01:05 AM	0	3	2	127	334	754
01:10 AM	0	2	18	127	316	752
01:20 AM	4	1	18	123	298	751
01:25 AM	11	5	17	112	281	746
01:30 AM	4	11	29	108	252	735
01:35 AM	7	4	14	101	238	731
01:40 AM	6	12	21	95	217	719
01:41 AM	6	22	9	89	208	697
01:45 AM	4	1	8	85	200	696
01:50 AM	12	3	45	73	155	693
02:00 AM	10	12	16	63	139	691
02:05 AM	3	5	12	60	127	686
02:15 AM	0	18	2	60	125	674

Deep Dive

Did Class or Gender Affect Survival?

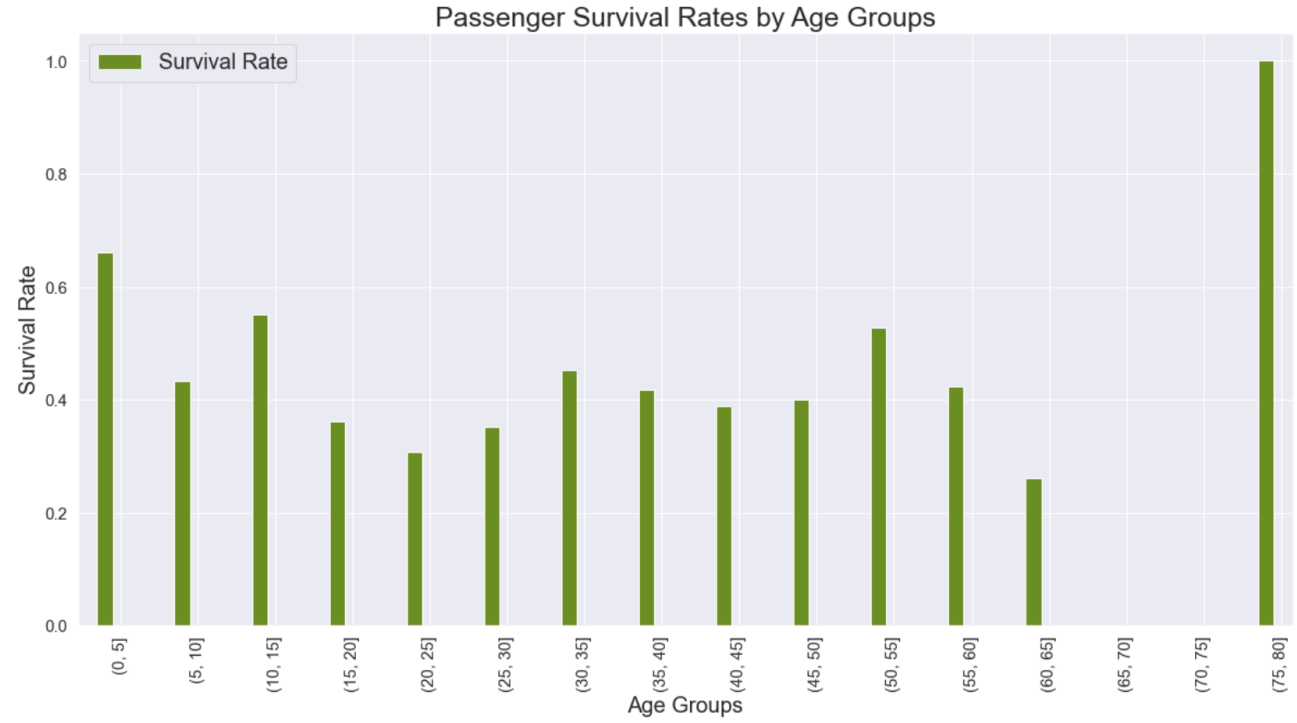
Objective: We began our analysis with the question of whether pclass was correlated with survival. In other words, we were curious to see if passengers of a certain pclass more likely to survive than passengers of another pclass. In order to obtain a more comprehensive understanding of our passengers, we divided the passengers into the following three categories: Women, Children, and Men. Passengers were deemed a child only if they were either a female under 18 or a male under 13 years of age.

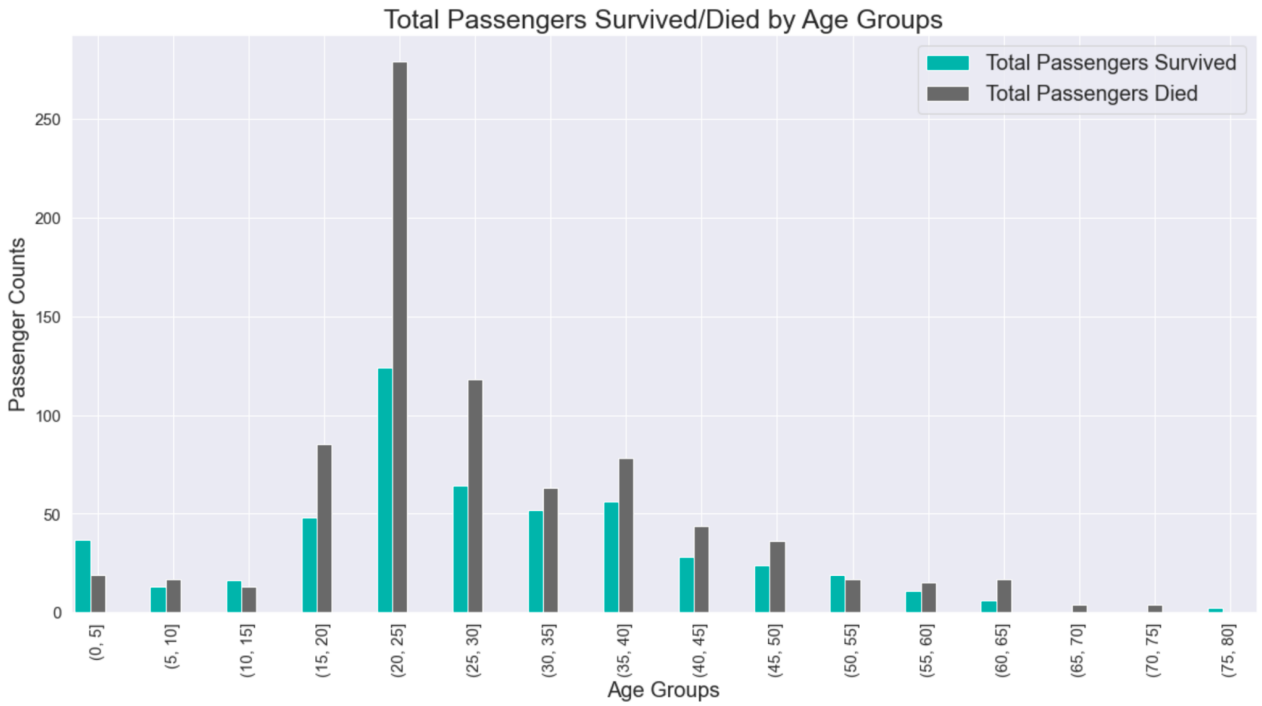


Insights: Per the graph above, children who embarked from pclass 2 had a 100% of survival. In other words, all children from pclass 2 survived. Children and women were more likely to survive if they were from pclass 1 and 2. Men were more likely to survive in pclass 1. Interestingly, men were more likely to survive if they were from pclass 3 than pclass 2. Overall, the graph reveals that men had the lowest percentages of survival regardless of pclass. With this insight, we decided to further explore the boat deployment timepoints to see if the rule “women and children first” held true for the Titanic. This analysis is explored a few sections later.

Did Age Affect Survival?

Objective: Next, we explored how age was correlated with survival rates. In other words, we were curious to see if passengers of a certain age group or set of age groups were more likely to survive than passengers of other age groups. We set age groups in increments of 5 and plotted the survival rate for each age group. In addition, we plotted the total survival and death counts by age group to understand the total distribution of passengers and their survival counts across age groups. We hypothesized that children were most likely to survive.

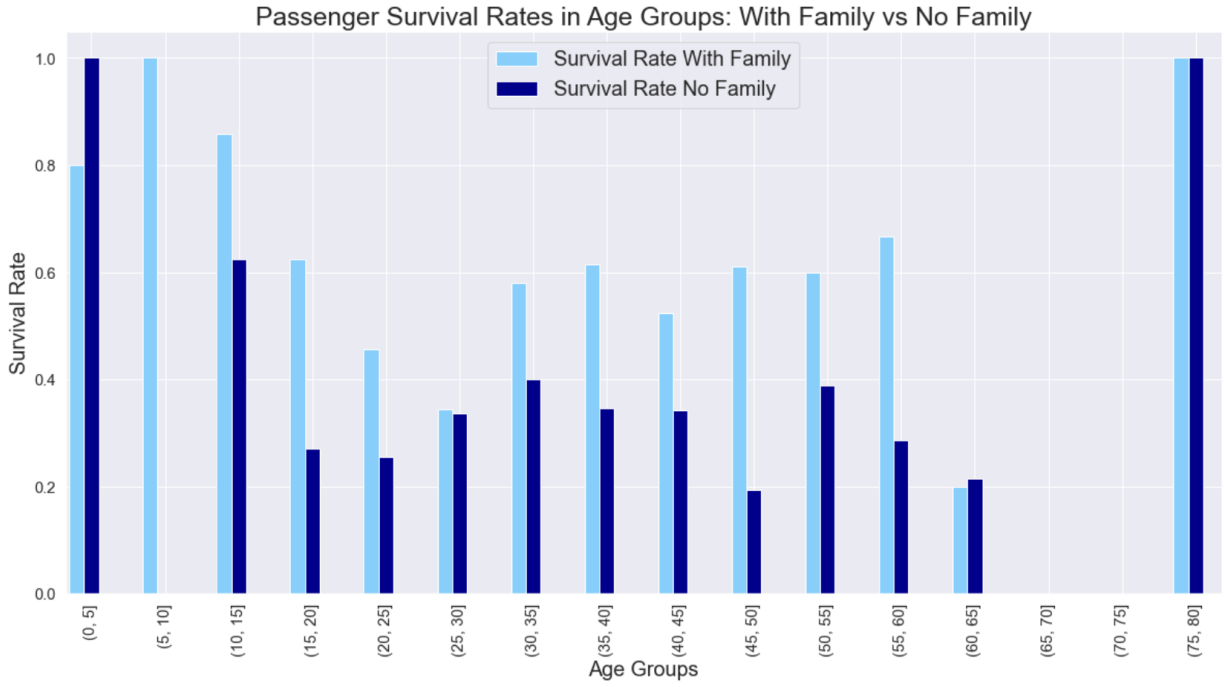




Insights: Surprisingly, those in the [75,80] age group were most likely to survive at a 100% chance of survival. When reviewing the second plot showing counts, we can see that there were only two passengers in this age group. After the [75,80] age group, young children in the [0,5] and [10,15] age groups were in fact most likely to survive. With these insights, we decided to dig deeper into survival rates, evaluating another facet of passengers: whether they had family or not.

Do Families that Stay Together Survive Together?

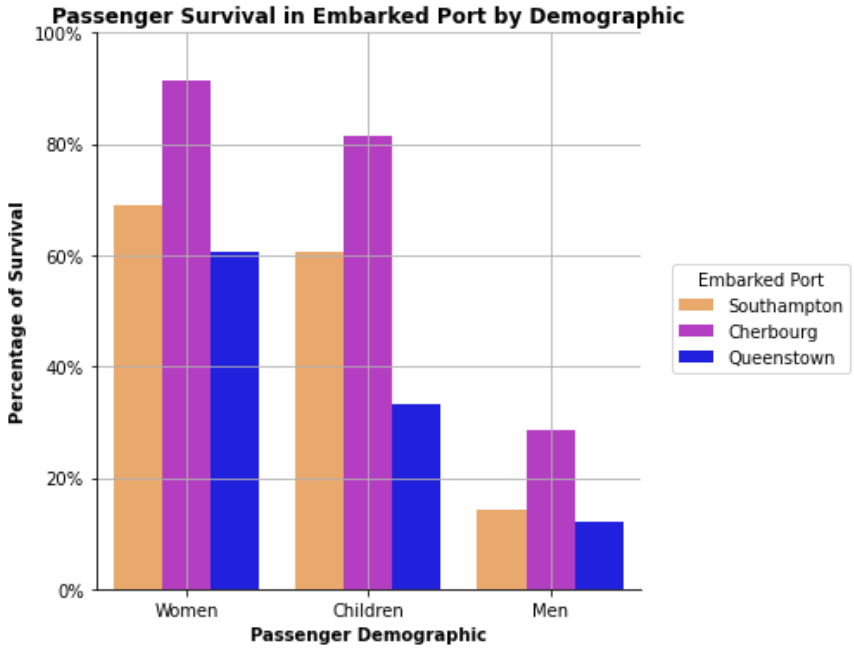
Objective: After reviewing the overall survival rates by age group, we considered family. Do families that stay together survive together? The Titanic dataframe provided had two numeric columns to denote family: sibsp (number of siblings or spouses), and parch (number of parents or children abroad), which we combined for this analysis. We hypothesized that those with family may have had a better chance of survival, as they could have alerted each other to the disaster, shared information for lifeboats, and advocated for each other to secure a spot on a lifeboat.



Insights: Having family aboard, in most cases, was related to a higher survival rate, with two exceptions: age group [0,5], and age group [60,65] in which those without family had a slightly higher survival rate. In age group [75,80], all passengers survived regardless of family. Those without family in the [45,50] age group had the worst rate of survival. In the age group [60,65], passengers had a poor rate of survival regardless of family.

Was Embarked Port Related to Survival?

Objective: The purpose of this section was to explore if a passenger's port of embarkation was related to survival. Similarly to the Pclass analysis above, we divided the passengers into the following three demographic groups: women, men, and children. We presumed that ports residing in affluent neighborhoods would attract more prestigious pclass 1 passengers while ports located in less-affluent locations would board more passengers into pclass 3. Our background research revealed that Queenstown, Ireland was home to poor and famished Irish immigrants recovering from the Great Famine in Ireland. Thus, we hypothesized that passengers embarking from Queenstown versus Southampton and Cherbourg would have the lowest percentage of survival.

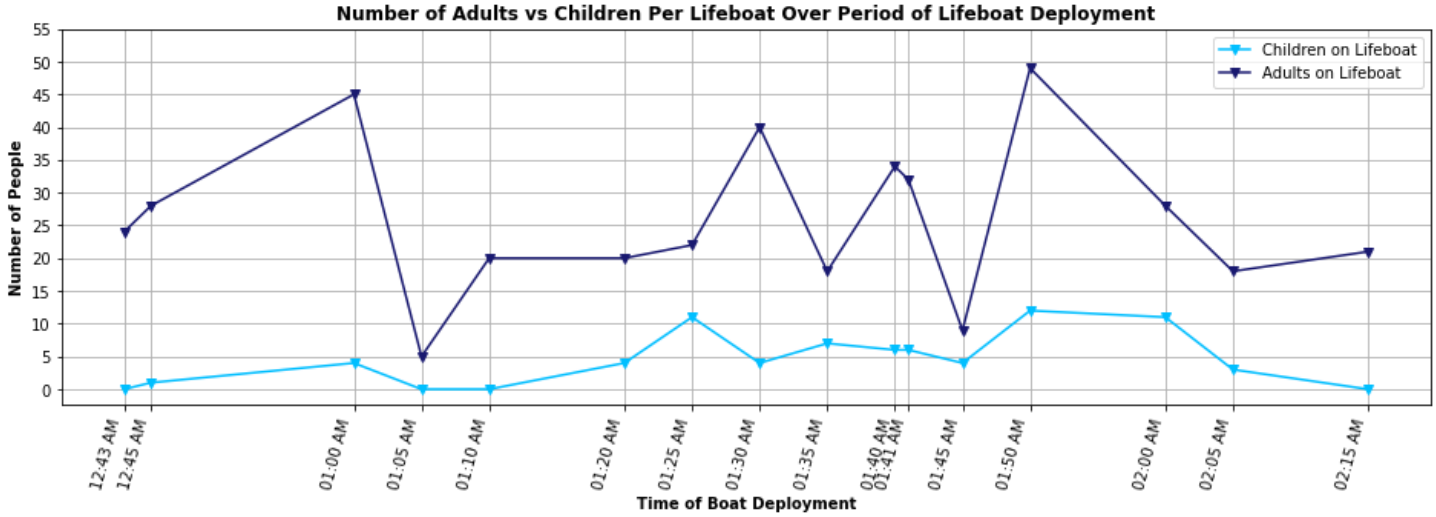


Insights: As depicted in the graph above, men who embarked from Queenstown had the lowest percentage of survival and women who embarked from Cherbourg, France had the highest percentage of survival. Overall, men had the lowest percentage of survival in Southampton, Cherbourg, and Queenstown.

Most notably, the graph reveals a similar pattern for each demographic group which is the highest percentage of survival from Cherbourg followed by Southampton, and the lowest percentage of survival from Queenstown. Thus, embarking from Cherbourg had the highest percentage of survival and embarking from Queenstown had the lowest percentage of survival.

Part 1: Women and Children First?

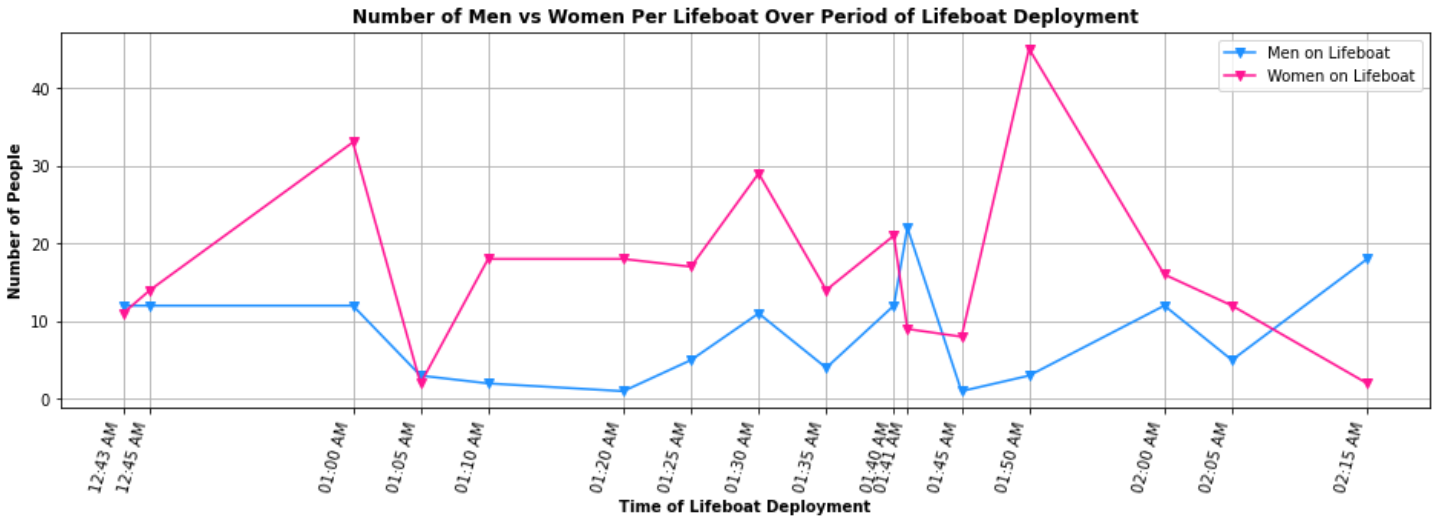
Objective: The origins of the maritime phrase "Women and Children First" dates back to the mid-1800s when the HMS Birkenhead sank due to striking rocks during its passage near the coast of South Africa. This mandate was also ordered by Titanic's captain, Edward Smith, at approximately 12:25 AM - 18 minutes before the first lifeboat's deployment. The order was intended to give women and children priority on the lifeboats over men. The objective of this first part was to see if the captain's orders held true for children on the first several lifeboats deployed from the Titanic. Our hypothesis was that his orders held true and thus, we would observe a larger number of children on the lifeboats that were deployed in the first, several lifeboats as opposed to the lifeboats deployed near the Titanic's sinking time at around 2:20 AM. We hypothesized that compliance towards the captain's orders would likely have been more readily accepted towards the initial boat deployment times rather than final stages near the Titanic's sinking due to panic and impending doom.



Insights: Per the graph above, the number of children began to steadily increase with the lifeboat deployed at 1:20 AM then plateaued from 1:30 AM to 1:45 AM. But most notably, four out of the total sixteen lifeboats contained no children. Of these four lifeboats with no children, three of them were amongst the first five lifeboats that were deployed at the following times: 12:43 AM, 1:05 AM, and 1:10 AM. The second lifeboat deployed at 12:45 AM only aboarded 1 child versus the 28 adults. This suggests that children did not have priority in the initial boat deployment timepoints.

Part 2: Women and Children First?

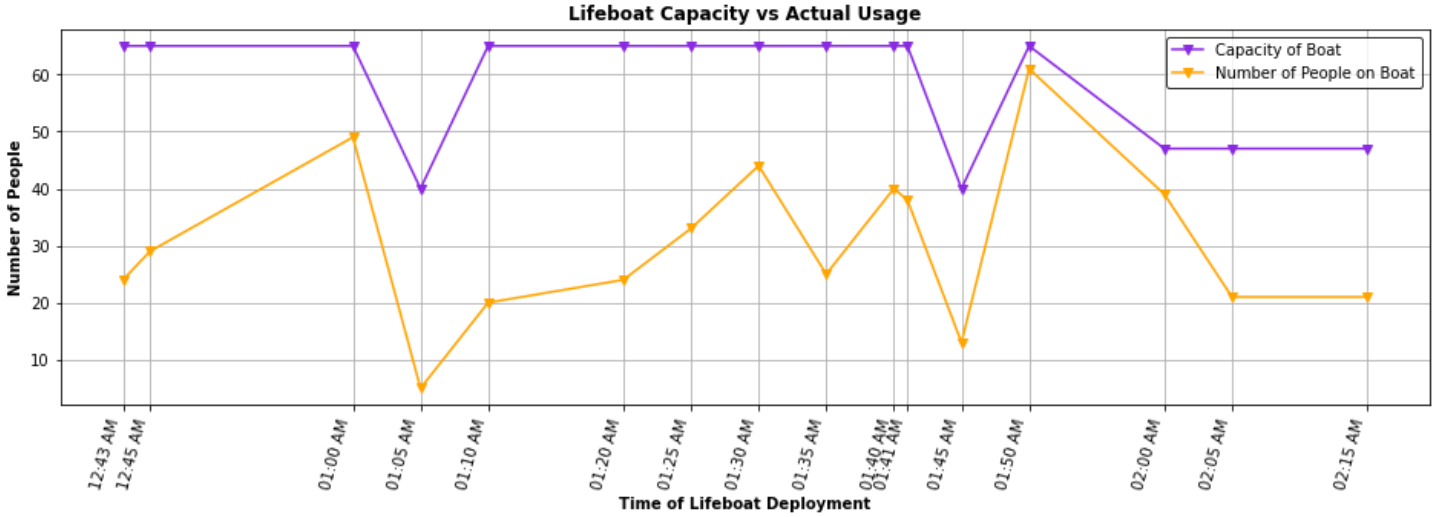
Objective: Part 2 of the analysis on "Women and Children First" was to observe if women had priority over men on Titanic's lifeboats. Similarly to Part 1, we hypothesized that the Captain's orders held true. Thus, we predicted our analysis would show a larger number of women than men on most lifeboats and greater compliance in the earlier stages of the boat deployment timepoints.



Insights: Per the time series graph above, 12 of the 16 boat deployment time points contained more women than men. This translates to about 75% of the lifeboats exhibiting compliance towards the Captain's orders. Most notably, we'd like to draw your attention to the last several lifeboats that were deployed. With the exception of the very last collapsible lifeboat deployed at 2:15 AM, the last several lifeboats deployed from 1:45 AM to 2:05 AM contained more women than men. This suggests that the passengers were compliant with the captain's orders even with the impending doom ahead. Of note, the very last collapsible lifeboat mainly consisted of men, because this lifeboat was never properly launched into the sea. During the process of its release from storage, it crashed onto the boat's deck, shortly before the sea washed over this particular area. Many male crew members who were working on overturning the boat managed to survive by climbing on this boat as it was washed out to sea.

Did Lifeboats Deploy Under Capacity?

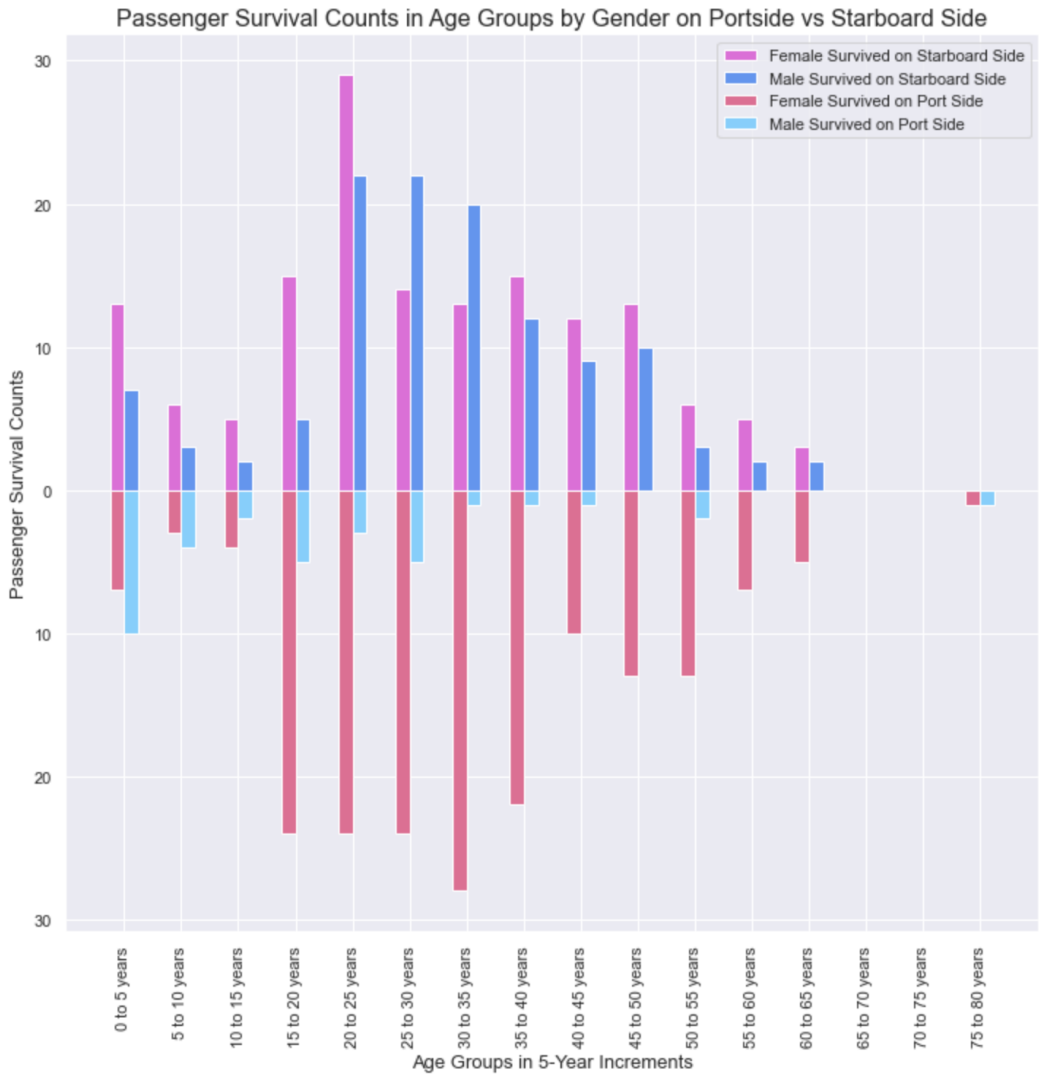
Objective: This last time series plot was to further explore if panic and impending doom affects usage of the lifeboats. We predicted that actual usage of the lifeboats would steadily increase and surpass the capacity towards the final stages of the boat deployment timepoints due to the effects of panic as Titanic began to perceptibly tilt port-side, the number of remaining lifeboats decreased, and more rockets were fired as serious distress signals.



Insights: Per the graph above, the plot shows a gradual increase in usage with the boats deployed from 1:10 AM to 1:30 AM. However, there is no overall increase in usage and none of the lifeboats ever pass its capacity even with the passage of time. Most lifeboats had a capacity of 65 along with 2 smaller wooden cutters that could hold 40 people and 3 non-collapsible lifeboats with a capacity of 47 that were launched towards the end of the boat deployment process. The two lifeboats both deployed at 1:50 AM, almost reached its capacity of 65 people. These two lifeboats were the last non-collapsible lifeboats that were launched.

The Titanic Sank With a Port Tilt: Did Boat Side Matter For Survival?

Objective: After the Titanic impacted on the starboard side, it developed a permanent tilt toward the port side (Vid). As such, we hypothesized that the port side may have flooded faster, and more people overall would survive leaving from the starboard side. Due to the limitations of cabin data (77%) missing, we were only able to analyze survivor data, based on which side of the ship their lifeboat launched from.



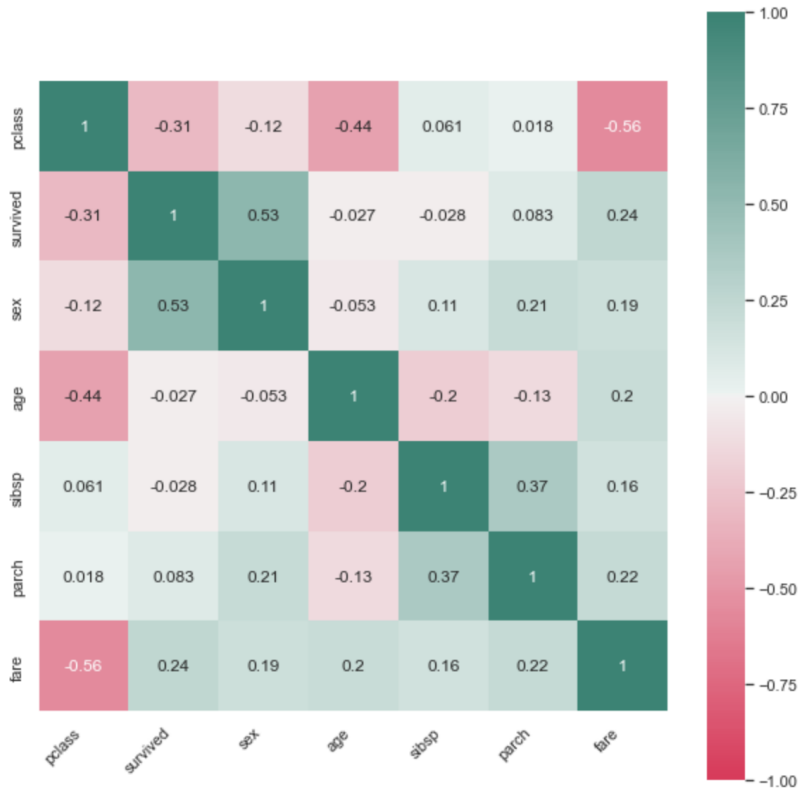
Insights: After plotting a quick bar plot to see how starboard survivor counts compared with port survivor counts, we found that more people in fact survived on starboard-side lifeboats (256 survived via starboard vs 220 via port). We decided to dig into these populations more by investigating survivor traits of gender and age groups. Female passengers survived in great numbers on both sides of the boat but skewed port side (138 starboard vs 177 port). Children, defined as under 18 for females, and under 13 for males, survived well on both starboard and port. Male passengers on the other hand, overwhelmingly survived on the starboard side (118 starboard vs 43 port), suggesting that the 'women and children first' rule may not have been implemented equally on both sides of the boat. We researched online to understand how the 'women and children' rule was implemented. Historic accounts show that on the port side, the women and children first rule was followed, allowing only a few men on each boat to row, whereas on the starboard side, men were allowed in empty seats.

Correlation Heatmap

Objective: After having completed our analyses we decided to use a correlation heatmap as a secondary check on previous insights, and also to explore potential other correlation insights between columns.

- Having changed the sex column to 'female' = 1, and 'male' = 0, we expected a positive correlation between sex and survival.
- We expected a negative correlation between pclass and fare (unintuitive because pclass 1 is socioeconomically the highest level).

- Earlier, having identified that those in a numerically lower pclass were more likely to survive, we expected a negative correlation between plass and survival, and a positive correlation between fare and survival.



Insights:

- Our expectations were met for:
  - A positive correlation between sex and survival of 0.53.
  - A negative correlation between pclass and fare of -0.56.
  - A negative correlation between pclass and survival of -0.31.
- This suggests that passengers who were in a numerically lower, higher prestige pclass, who paid more, were more likely to survive. We hope that in disasters, people of different classes would be treated equally, but unfortunately the data suggests that was not the case for passengers of the Titanic.
- We noticed a correlation of 0.37 for parch and sibsp, meaning if a passenger had parents/children, it was more likely that they also had a sibling or spouse on board. In addition, those with family are slightly more likely to have been a woman.
- We also noticed a negative correlation for both pclass and age at -0.44, meaning those who had a prestigious pclass were more likely to be older than those who had a less prestigious pclass.

Conclusion

In conclusion, the factors that most affected survival include:

- **Sex and Age:** Women and children had the highest chances of survival, while men had the worst chances.
- **Family:** Passengers who had family on board were more likely to survive than passengers who traveled without family.
- **Embarked Port:** Women, children, and men who embarked from Cherbourg had the highest percentage of survival.
- **Boat Side:** Male passengers who attempted to board a lifeboat on the port side of the Titanic were more likely to have been barred from doing so.

We tested phrases that people say about the Titanic disaster and found:

- **“Women and Children First”:** Children did not have priority in the first 5 lifeboats that were deployed. Exactly 75% of the boat deployment timepoints had more female than male passengers.
- **“Lifeboats were spent unused or underutilized”:** There was no overall increase in usage and none of the lifeboats ever passed its capacity even as the Titanic approached sinking.

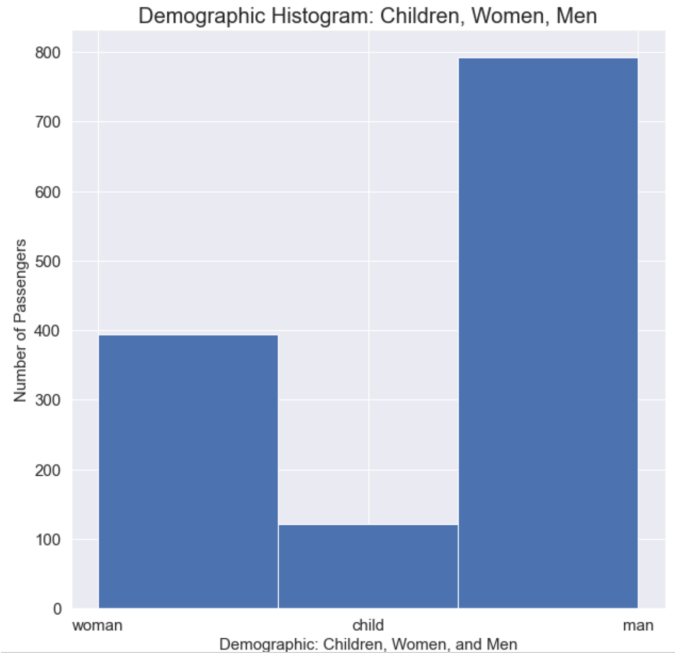
Best chance primary characteristics: Children in pclass 2, Women in pclass 1, Children in pclass 1
Best chance secondary characteristics: Embarked from Cherbourg
Worst-chance primary characteristics: Men in pclass 2
Worst chance secondary characteristics: Embarked from Queenstown who attempted to board a lifeboat on the port side of the Titanic.

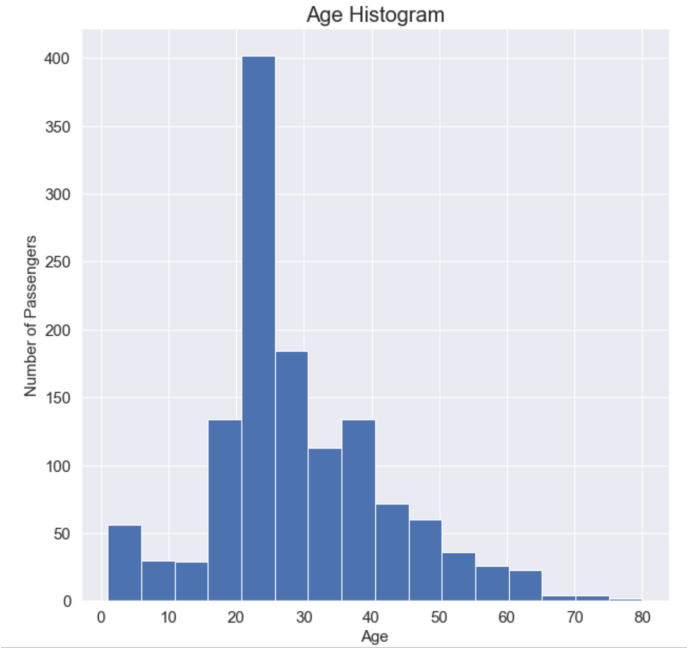
Best chance primary survival characteristics are a subjective assessment based on our analyses using passenger demographic (women, children, or men), and pclass (1, 2, or 3). Secondary characteristics were based on inference from our analyses due to the limitations of our data. These include: embarked port, and port side vs starboard side.

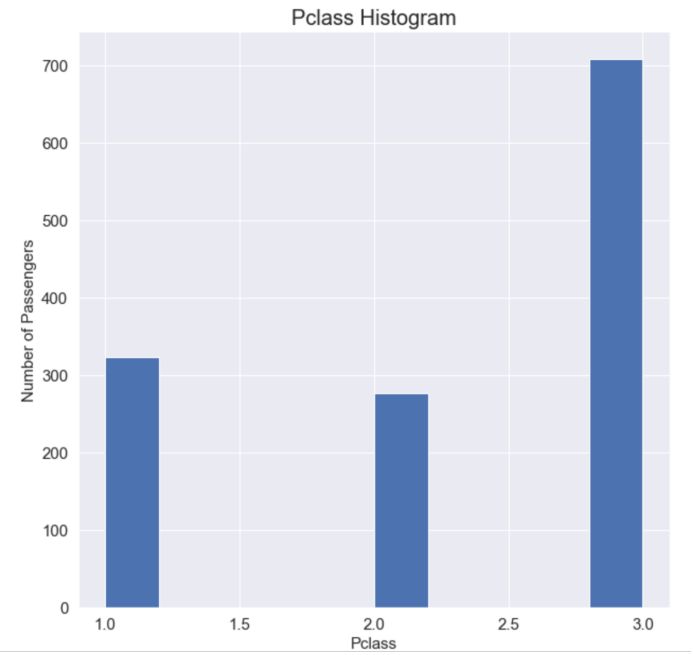
Appendix A: Null Counts

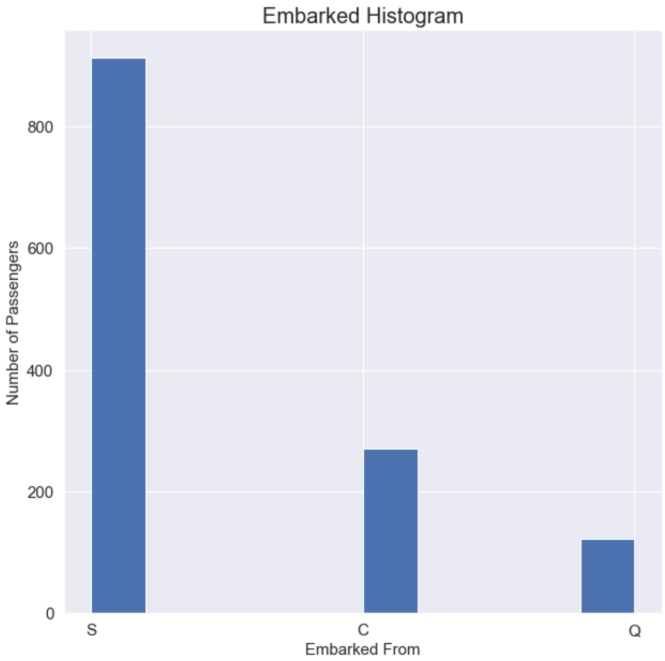
Null Counts		Non-Null Counts		Dtypes	
pclass	0	pclass	1309	pclass	int64
survived	0	survived	1309	survived	int64
name	0	name	1309	name	object
sex	0	sex	1309	sex	object
age	263	age	1046	age	float64
sibsp	0	sibsp	1309	sibsp	int64
parch	0	parch	1309	parch	int64
ticket	0	ticket	1309	ticket	object
fare	1	fare	1308	fare	float64
cabin	1014	cabin	295	cabin	object
embarked	2	embarked	1307	embarked	object
boat	823	boat	486	boat	object
body	1188	body	121	body	float64
home.dest	564	home.dest	745	home.dest	object
boat_loc	833	boat_loc	476	boat_loc	object
boat_deploy_time	833	boat_deploy_time	476	boat_deploy_time	object
capacity	833	capacity	476	capacity	float64
dtype: int64)		dtype: int64,		dtype: object	

Appendix B: Basic Histograms






Pclass Histogram
Number of Passengers
0
100
200
300
400
500
600
700
1.0
1.5
2.0
2.5
3.0
Pclass


Embarked Histogram
Number of Passengers
0
200
400
600
800
S
C
Q
Embarked From