Analysis of “Titanic 1912” Passengers

Trang Tran

ALY6000: Introduction to Analytics

Feb 11, 2023

Overview of the event and the dataset

The dataset demonstrates information on 891 passengers (not all passengers) in the sinking event of RSM Titanic 1912, as noted in the Data Dictionary beside.

Firstly, I did the data cleaning process to remove all “No data” records, handle other inappropriate values and select the analyzable variables for my analysis.

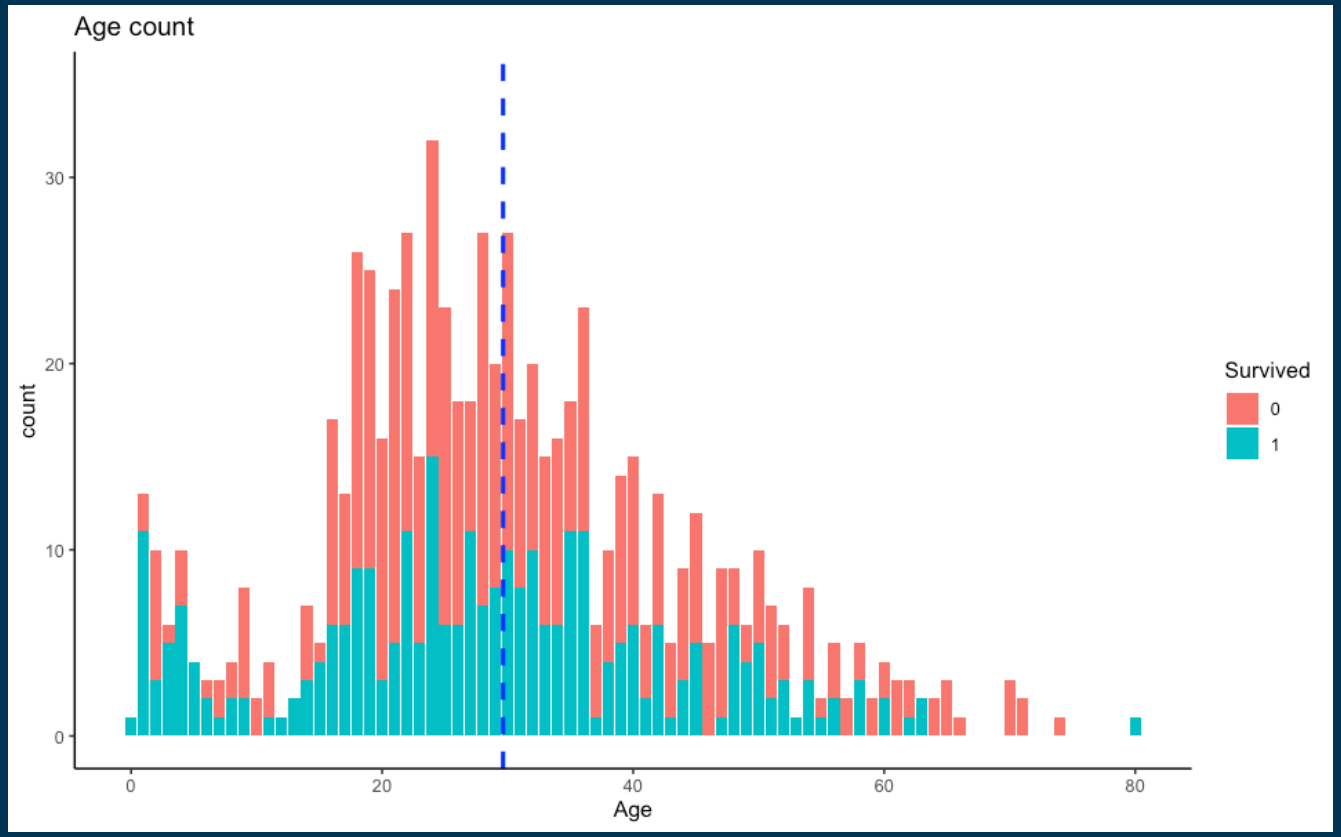
Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked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

What is the distribution of age in the Titanic passengers vs. the mean age?

This figure shows the mean age of ~30 and the high population of the age range of 15 to 35 (labor age) in the total passengers.

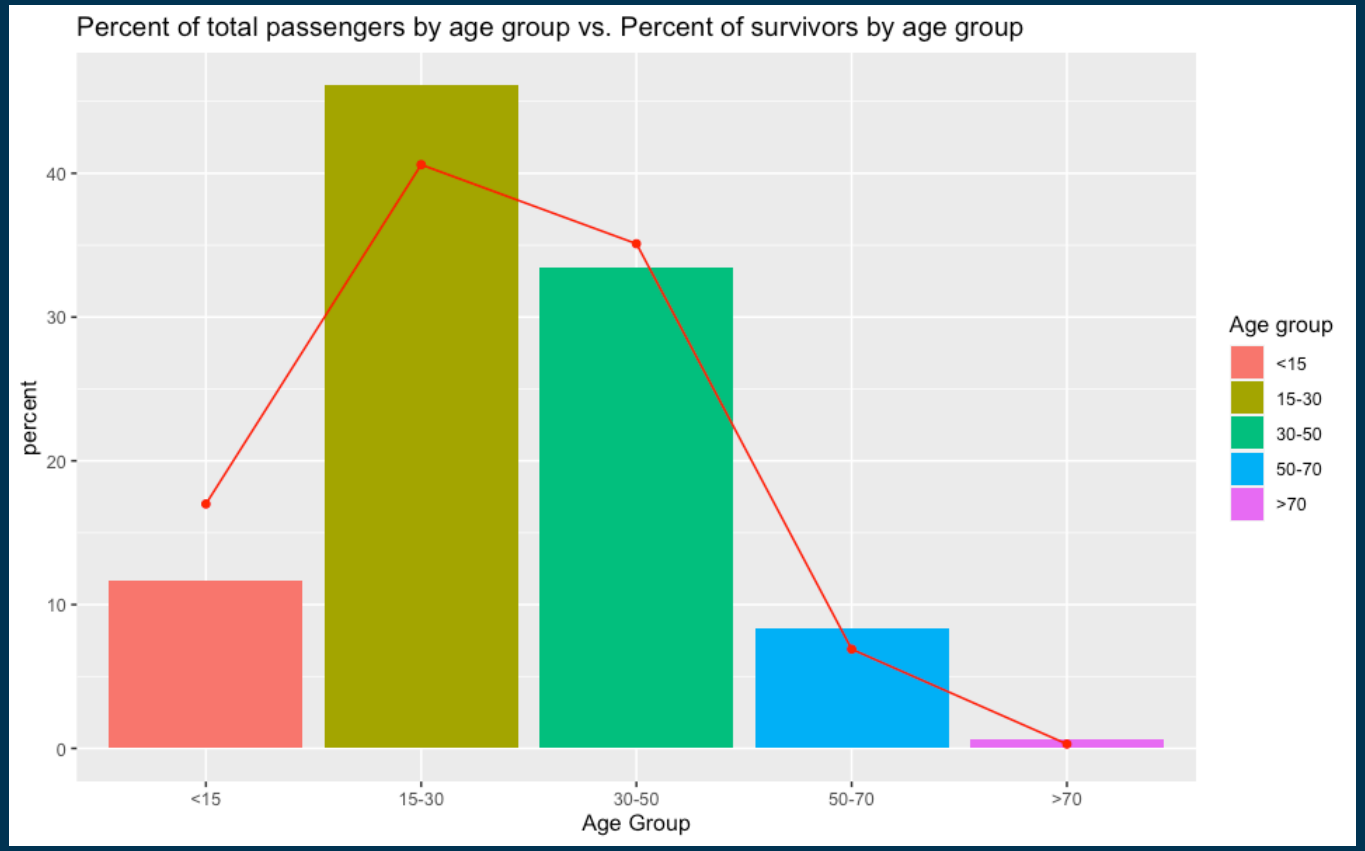
I want to better understand the percentage of age groups and the survival rate of these groups. Therefore, I did a group descriptive as follows:

gr_by_age	count	percent	count_sv	sv_pct
<15	83	11.7	49	17
15-30	328	46.1	117	40.6
30-50	238	33.4	101	35.1
50-70	59	8.3	20	6.9
>70	4	0.6	1	0.3

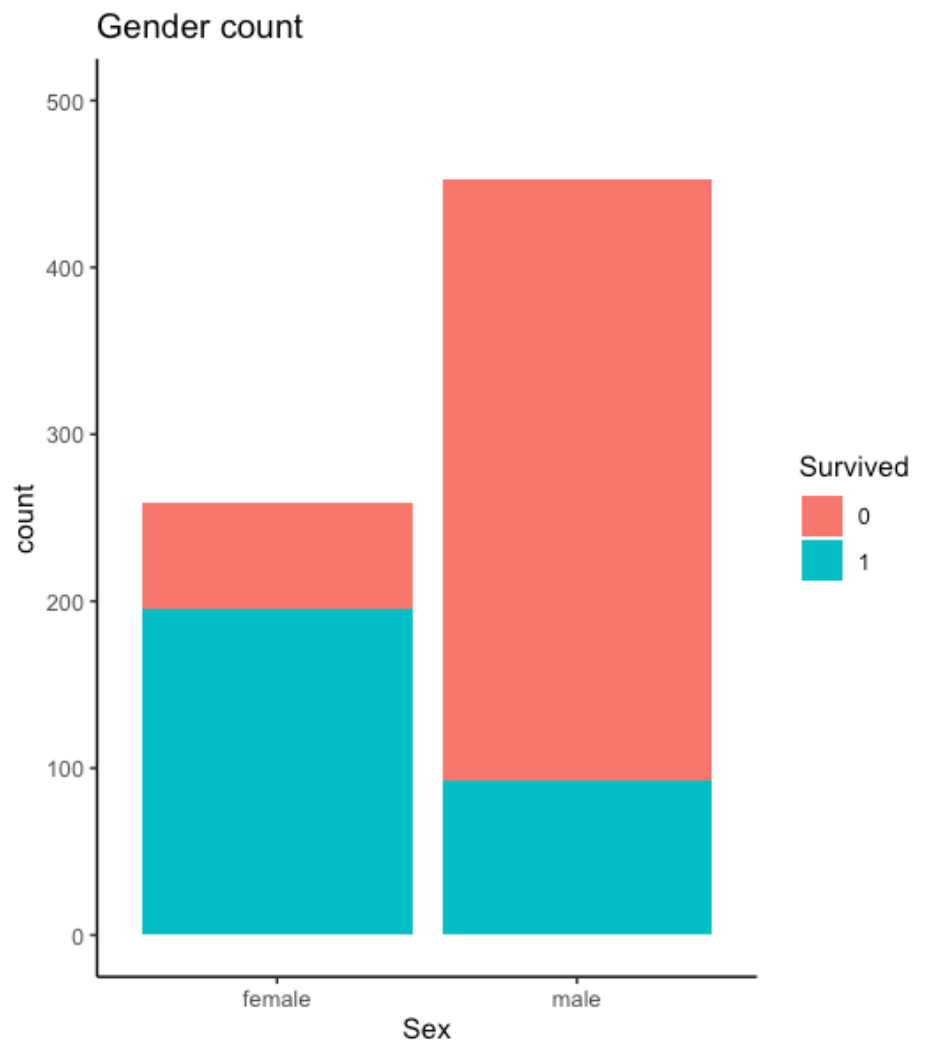


The percent of total passengers vs. the percent of survivors by age groups

The column visualization represents the percentage of age groups' shares in total Titanic passengers, while the red line graph shows the percentage of survivors divided by these five age groups. Overall, there is no big relative difference between the two percentage indicators. The survival rate of the age group below 15 has slightly higher than its percentage per total people.

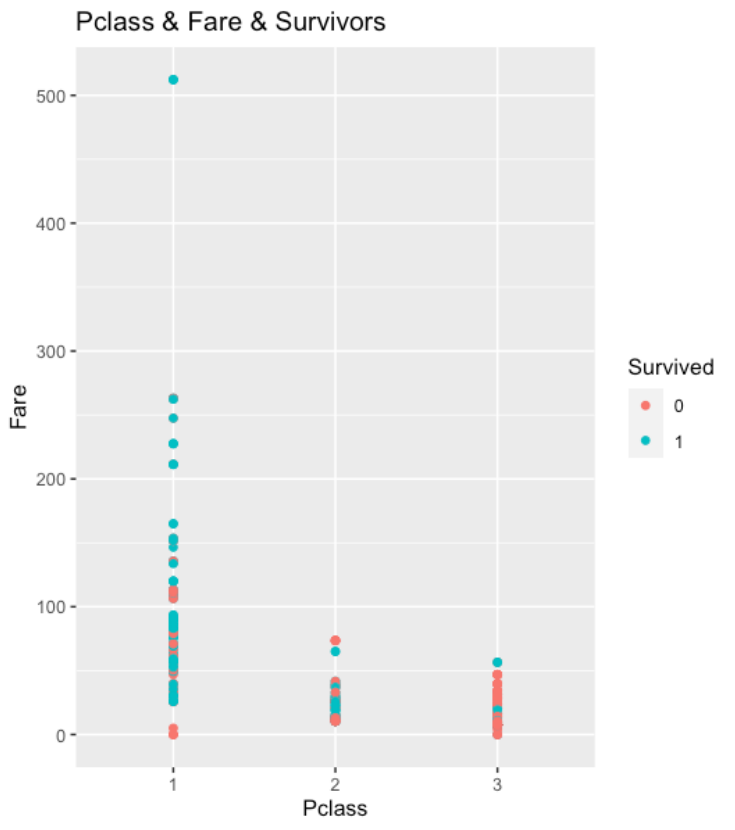


Whether women were given priority in escaping at that event?

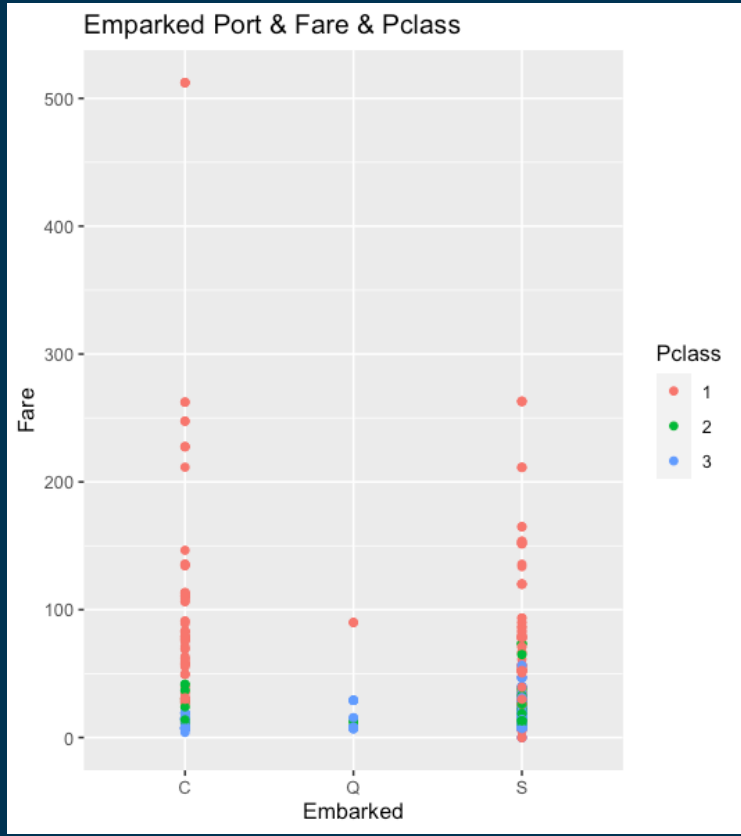


The number of males boarding this ship was much higher than the number of females. Yet, the female survivors doubled the number of male survivors. There must have been a priority for women at this tragic event.

The distribution of passengers in terms of ticket class, fare, and port of embarkation

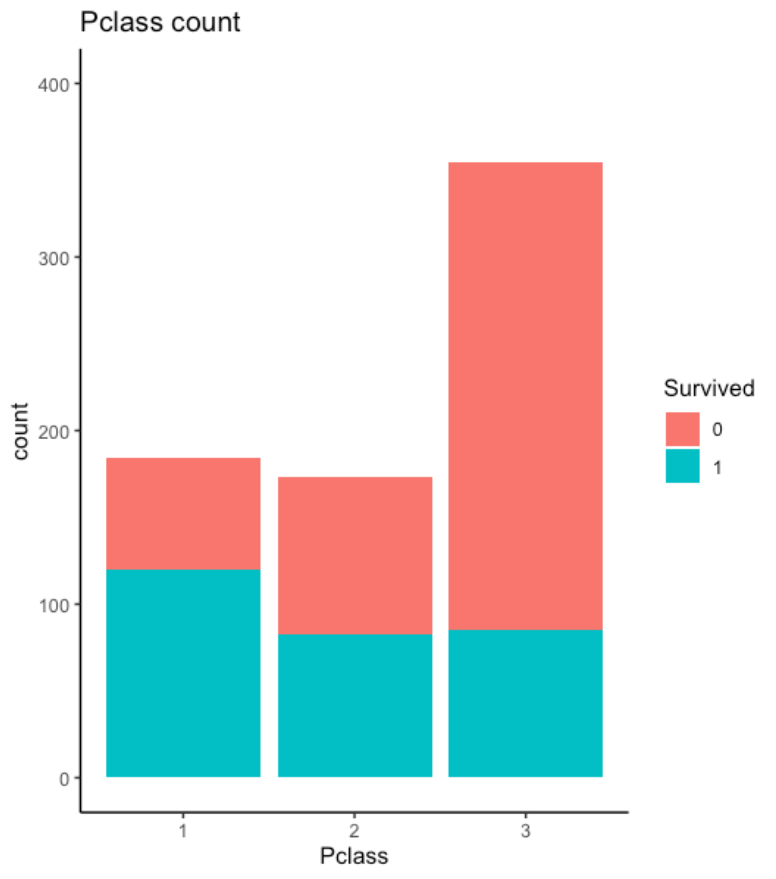


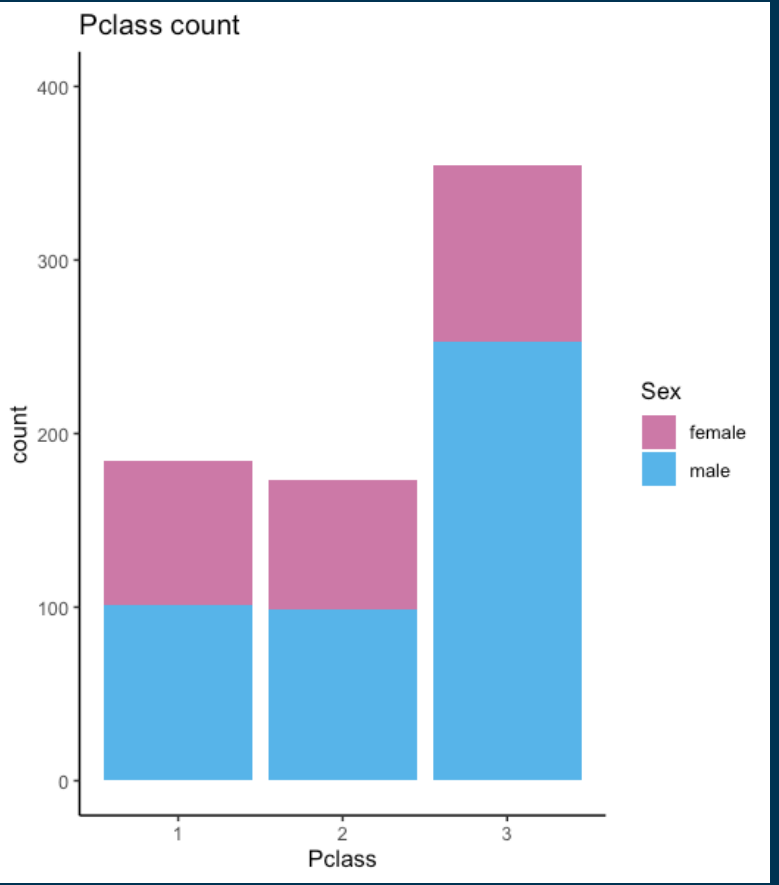
Pclass 1 had the highest fares, while P2 & P3 had fares below 80 and were quite similar in value.



Most Pclass 1 passengers embarked on the C & S ports only, while the other two P-classes were divided into all three ports.

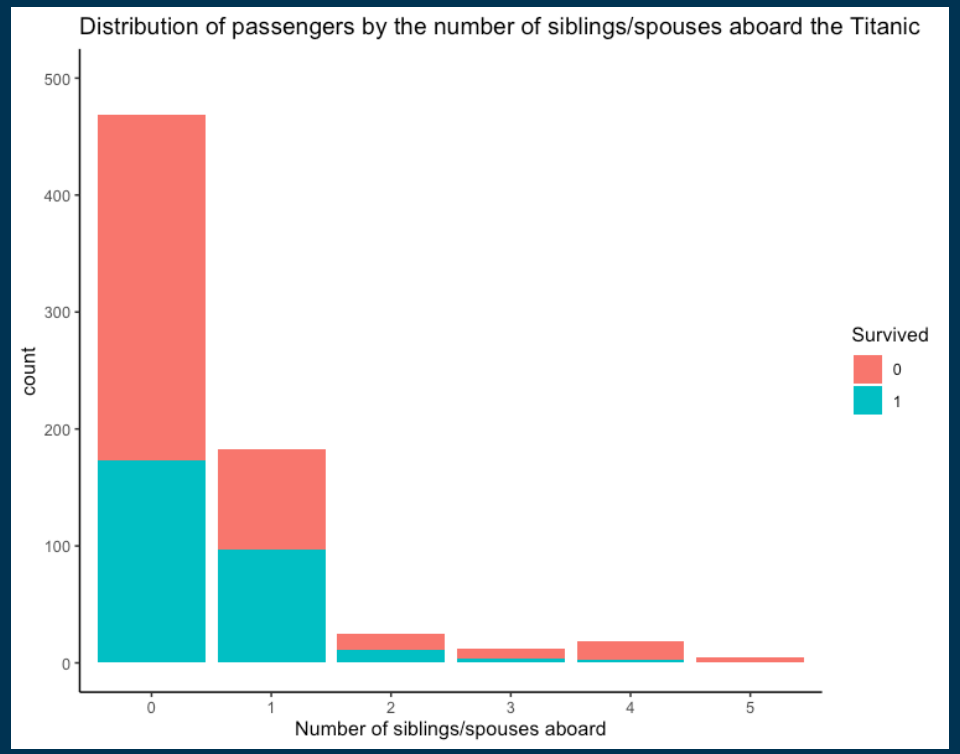
The ticket class count, survival rate, and gender share

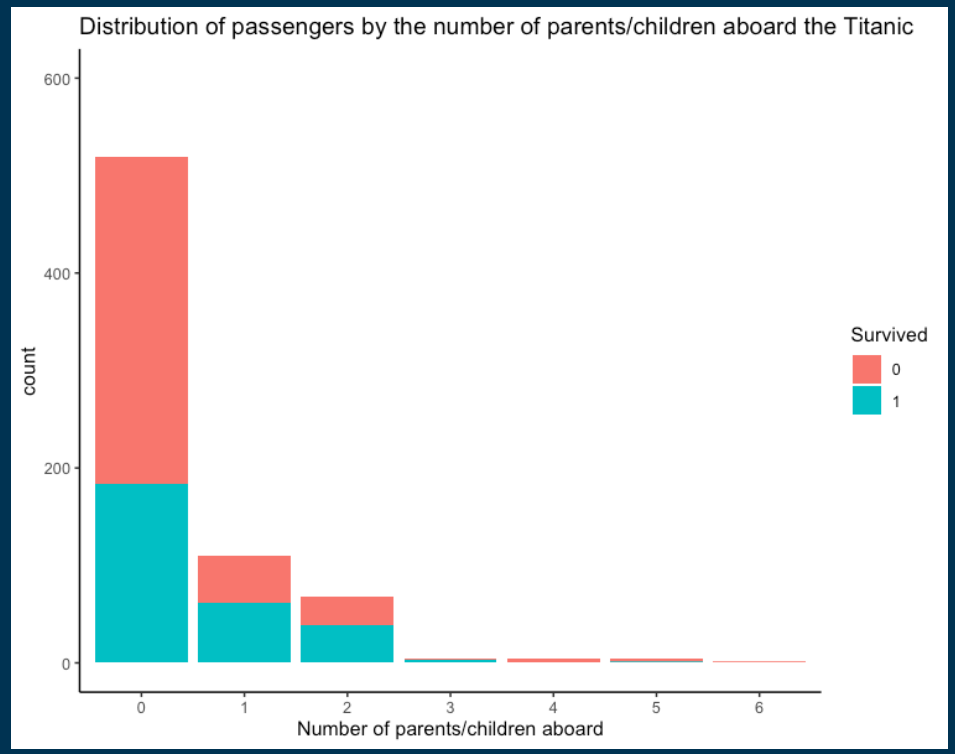




The number of 3rd class passengers was approximately equal to the sum of the other two P-classes, while the survival number of this class was similar to P2 survivors, even lower than the P1 survivors. Were the P1 & P2 prioritized? We cannot conclude, because the women's share of the P1 & P2 groups was high (almost 50% of each group), but the women’s share of the P3 group was just around ¼. Women and children are always the highest priority, no matter what ticket class.

The distribution of passengers by Siblings/Spouses and Parents/children variables





The number of people traveling alone (not traveling with family) overwhelmed per total number of Titanic guests.

References

ProgrammingR. Data Cleanup: Remove NA rows in R. Retrieved Feb 5, 2023.

https://www.programmingr.com/examples/remove-na-rows-in-r/

dplyr 1.1.0. Mutate multiple columns. Retrieved Feb 5, 2023.

https://dplyr.tidyverse.org/reference/mutate_all.html

Kaggle. Titanic Dataset. Retrieved Feb 4, 2023.

https://www.kaggle.com/datasets/yasserh/titanic-dataset?datasetId=1818188&language=R

Stackoverflow. How to change legend title in ggplot. Retrieved Feb 10, 2023.

https://stackoverflow.com/questions/14622421/how-to-change-legend-title-in-ggplot

Stackoverflow. How do I manually set geom bar fill color in ggplot. Retrieved Feb 10, 2023.

https://stackoverflow.com/questions/18229835/how-do-i-manually-set-geom-bar-fill-color-in-ggplot

Statistics Globe. R ggplot2 Error: Continuous value supplied to discrete scale (2 Examples). Retrieved Feb 10, 2023.

https://statisticsglobe.com/r-error-continuous-value-supplied-to-discrete-scale

Other R cheat sheets