Titanic Dataset - Group Project Written Report

Puja Shah, Madeleine Moon, Jonah Engelmann, Doreen Chang
Department of Information Systems and Analytics: Santa Clara University

Professor Tan
OMIS 112: Data Visualization
June 8, 2025

Introduction and Background

After some team members recently saw the movie *Titanic*, we were interested in the details of the real-life journey that led to the maritime disaster. Sharing this information we recently learned, with the rest of the group, all of us became interested in the story of the Titanic and did our own research. We learned there were many different kinds of data recorded from the ship's journey, that we could connect Tableau's features with the Titanic dataset that we used for this project, to understand how these different data - like age, gender, ticket class, and embarkation point - came together and if they could give us more information about how they influenced survival on this journey and more broadly human behavior under crisis, to offer a deeper insight into our new interest in the Titanic. Specifically, we wanted to answer questions:

- What factors most influenced survival rates?
- How did class, gender, age, and family size affect outcomes?
- Are there geographic patterns in passenger embarkation and survival?

Our group felt that this would have broader benefits to others, like providing insights to data storytelling, social science, and reaction in crisis times, and consequently how to create crisis response strategies. We felt it could be valuable to data & behavioral analysts & historians.

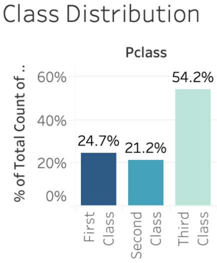
Datasets

From Kaggle's Titanic Dataset we used, "Titanic EDA with Tableau." The dataset includes 14 variables including, "pclass, survived, name, sex, age, fare home.dest" and a few others. Pclass refers to the class of the passenger - from 1st to 3rd classes. Survived indicates whether the passenger lived, represented by a 1, or died, represented by a 0. Name refers to the passenger's name on the ticket, sex, if the passenger was female or male, age for passenger age. Fare referred to the British Pounds the rider paid for their ticket. (*Titanic dataset*).

The dataset had 1,309 rows of passengers' data cases. The locations in our dataset were pertaining to "embarked" or the city that they boarded the Titanic - either Southampton, Cherbourg, and Queenstown. As well as pertaining to "home.dest" or the City and Country/State that the person was born from. These variables helped us to reach our goal since we were made statistical correlations with survival to be analyzed across the demographics. Some limitations we identified were some missing values in variables like age. Home.dest was particularly messy. Most rows had no values or sometimes the city and other times the country were listed. So, we created a second dataset called titanic_location_mapping containing the city and country of each passenger along with the original home.dest column to allow us to join the datasets together. We'd also liked to see more data on the passengers' intentions or physical conditions as we felt that would've affected survival.

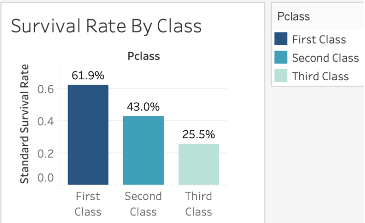
Data Story

To identify which factors most influenced survival during the Titanic, we created a series of visualizations in Tableau analyzing passenger data by class, sex, and age.

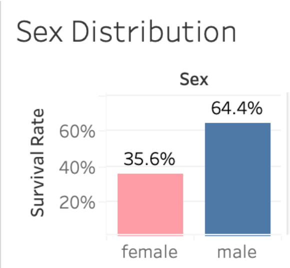


Distribution and Survival Rate of Class, Sex, and Age Bar Charts: First, we constructed a class distribution bar chart by dragging Pclass into Columns and CNT(Pclass) into Rows. We also placed Pclass under Color to differentiate the classes and Rows under Label to show each class's percentages. This revealed that 3rd class passengers made up ship majority at 54.2%, while 1st class passengers accounted for 24.7% and 2nd class passengers accounted for 21.2%. This distribution reflects the typical composition of transatlantic travel during this era, with most individuals being working-class seeking new opportunities in America.

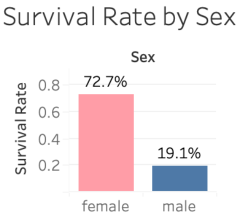
To assess the impact of class on survival, we created a calculated field called "Standard Survival Rate" using the formula AVG([Survived]) to produce the percent of passengers who died. We then built a chart measuring the survival rates for each class by placing Pclass in Columns, ACG(Standard Survival Rate) in Rows, Pclass under Color, and ACG(Standard Survival Rate) under Labell. The results showed that 61.9% of first-class passengers survived, compared to 43% of second-class passengers and just 25.5% of third-class passengers. This sharp decline in survival rates suggests that higher class passengers were more likely to survive, potentially due to their proximity to lifeboats or preferential treatment during evacuation.



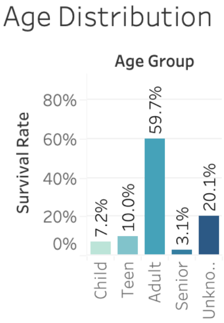
Next, we examined the role of gender by creating a sex distribution chart. Using the same construction method as the class chart, we placed Sex in Columns and CNT(Sex) in Rows, adding Sex under Color and CNT(Sex) under Label for clarity. This chart showed that the voyage was male-dominated, with men comprising 64.4% of passengers and women comprising 35.6% of them. It is important to note, however, that this stark gender imbalance was typical for the time, as men would often travel ahead of their families.



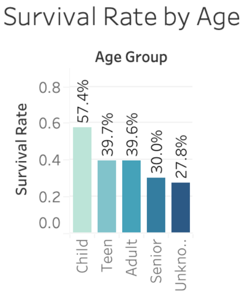
To understand how gender influenced survival, we constructed a corresponding bar chart dragging Sex into Columns and ACG(Standard Survival Rate) into Rows, with Sex under Color and ACG(Standard Survival Rate) under Label. The survival rates by gender revealed one of the most dramatic disparities in our dataset. While 72.7% of women survived, only 19.1% of men did. This nearly four-to-one difference in survival rates reflects the "women and children first" policy that was strictly enforced during the Titanic evacuation.



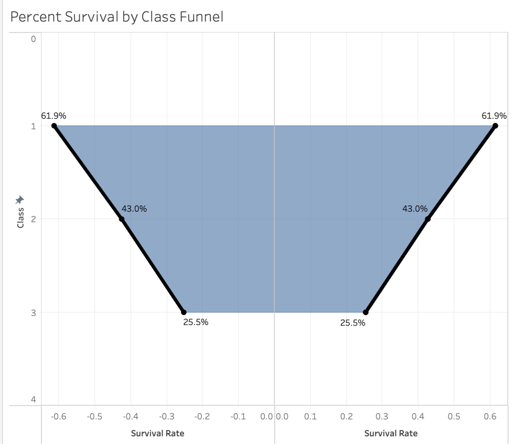
As a final measure, we decided to look at the correlation between age and survival rate. Before building our age distribution chart, we created a calculated field called "Age Group" to categorize passengers as children (under 13), teens (13-19), adults (20-59), seniors (60 and older), or unknown (no age recorded). Then, we employed the same construction method as the s chrartspreviouscharts, replacing Age Group and CNT(Age Group) as the variables. The results showed that adults represented the largest age group at 59.7%, followed by teens at



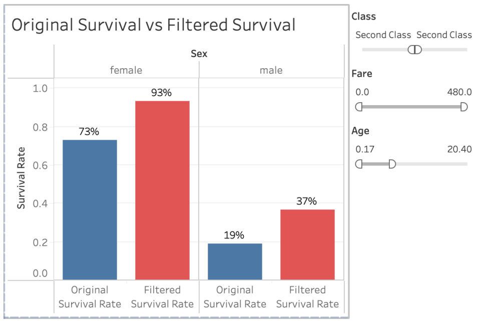
10%, children at 7.2%, and seniors at 3.1%. Notably, 20.1% of passengers had unknown ages, highlighting a significant gap in the historical data.



To create our final bar chart comparing survival rates, we placed Age Group in columns, ACG(Standard Survival Rate) in rows, Age Group under Color, and ACG(Standard Survival Rate) under Label. This chart revealed that 57.4% of children survived, with teens trailing behind at 39.7%, adults at 39.6%, seniors at 30%, and those with unknown ages at 27.8%. This finding, once again, reinforces the strict implementation of the "women and children" first policy during the Titanic evacuation.

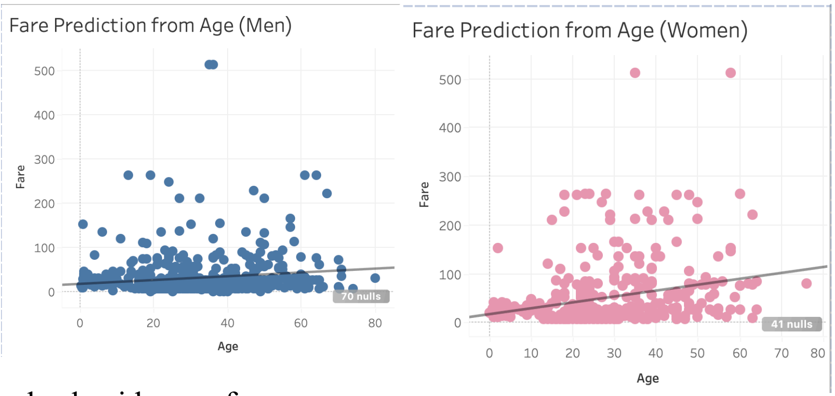


Percent Survival by Class Funnel Chart: We wanted an alternate, advanced view of the percent survival for each class and how it varies between the passenger classes. We created a calculated field, Total Passengers Per Class, where we used an LOD expression - { FIXED [Pclass] : SUM(1) }.We made another calculated field, % Survived per Class. We then created a duplicate negative version of this field, Negative % Survived per Class, to produce the other half reflection of the funnel. We then made & put calculated field, Negative pClass, into rows and made it descending and put % Survived per Class and Negative % Survived per Class in columns and created a dual axis. We set labels marks to see the % of survival in each class's number. We set the y-axis for Class to have a custom fixed start from 0 and end to 4 and a reversed scale, to see the changes starting from 1st to 3rd class. This helps address our goal of understanding how class affected survival outcomes, as we can see that 1st Class passengers had a 61.9% survival rate, going down to 2nd Class having a 43% survival rate, and finally 3rd Class having a 25.5% survival rate.

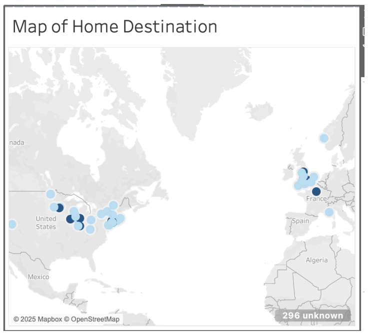


Class Fare and Age Multi Filter: We spent a lot of time looking at how class, fair, age, and gender affected survival individually, but we were also curious these factors affected survival at the same time. To examine this, we built a double bar chart showing women's global survival vs their survival with any combination of age, class, and fare filters. We did the same thing with men's survival. To make this double bar chart, we used a fixed LOD expression allowing the bar on the left to show the average survival rate for all women and men without the filters. This interactive feature allowed easy comparison between women/men on average, and women/men with a specific filter. For example we could compare womens overall survival rate with first class women over the age of 50 who paid over 200 pounds. There were many takeaways from this chart such as that young and old people in second class had much better survival odds than middle aged people in second class.

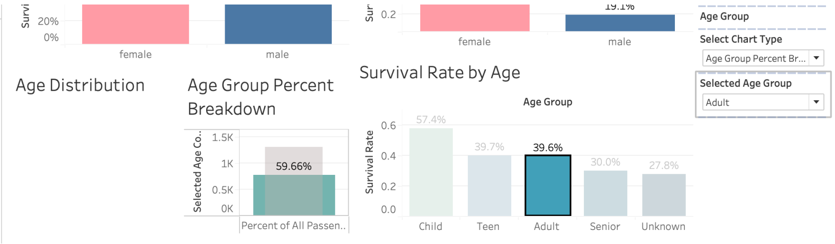
Fare Prediction from Age charts: We learned from an earlier chart that age had an influence on survival, so we wanted to examine age's relationship with fare: specifically if older passengers were paying more on average. Using a simple scatter plot with age on the s-axis and fare on the y-axis, we were able to examine this relationship by adding a trendline. For women, the trendline had a positive slope indicating that older women indeed paid more fare on average. This was also true for men but at a lesser degree.



Home Destination vs Survival Map: We were also very curious if passengers' home destination had any impact on survival. Questions like: "did the all european crew favor putting european passengers on the lifeboats?". The home destination data was very messy and required us to make a new dataset with the city and country only as mentioned earlier. Dragging the city, country, name, and survival columns to the detail shelf gave us a nice map of where people were from and if they survived. The results were clear: home destination had no real impact on survival.



Age Group Survival to Age Group Percentage Breakdown Interactive: While the Age Distribution chart shows the percentage of each age group in the boat, we were interested in visualizing how a specific age group's proportion, such as Child, looks like in comparison to the full 100% population of the boat. We added a button to select between the two chart types. If you select Age Group Percent Breakdown Chart then when you click on a specific age group of the Survival Rate by Age Chart, like Adult, then you can see the proportion of adults to total passengers or their 59.66% share of the total 100% of passengers on the Age Group Percent Breakdown Chart. We also created the Age Group Percent Breakdown Chart by creating a parameter titled Selected Age Group, where we added in string values of Child, Teen, Adult, Senior, and Unknown. We then made a calculated field titled Selected Age Count as well as another calculated field, Total Age Count; we placed both in rows as SUM values. We also created another calculated field titled One Bar Label and then we brought that field to columns. We created a dual axis and



Select Chart Type: Age Group Percent Br...

Selected Age Group: Adult

synchronized axis, as well as altered the colors so that the SUM(Total Age Count) was a transparent gray and made the size smaller and SUM(Selected Age Count) was green. We went to the dashboard and added the action to change parameter and a % of Total Label calculated field, which we added to the label marks card to show the percentages.

Summary and Conclusions

Completing this project has helped our group to understand various factors that influenced survival on the Titanic, especially gender, age, and class. We found that women, children, and first class passengers were more likely to survive. We were also able to use the charts to explore how factors like age and fare were correlated. This project pushed us to use our learnings from the class to create bar, map, funnel, and predicative charts with LOD expression and interactive features to help analyze and identify data patterns for data storytelling.

References

Parulpandey. (2020a, January 25). *Titanic Eda with tableau*. Kaggle. https://www.kaggle.com/code/parulpandey/titanic-eda-with-tableau

IMDb.com. (1997a, December 19). *Titanic*. IMDb. https://www.imdb.com/title/tt0120338/

Titanic dataset. Titanic Dataset - NannyML 0.10.4 documentation. (n.d.-a). https://nannyml.readthedocs.io/en/v0.10.4/datasets/titanic.html#:~:text=boat%20%2D%20lifeboar%20information%20if%20the,and%20destination%20information%2C%20if%20available.