Project 2: Visualization Ethics and Communication

By Group 7

# Group Members:

* Richard Hoehn (**richardhoehn**)
* Hector Rogel (**rogel9007)**
* Isaiah Osborne(**IMOsbo**)
* Karson Woods (**Kwoods132**)

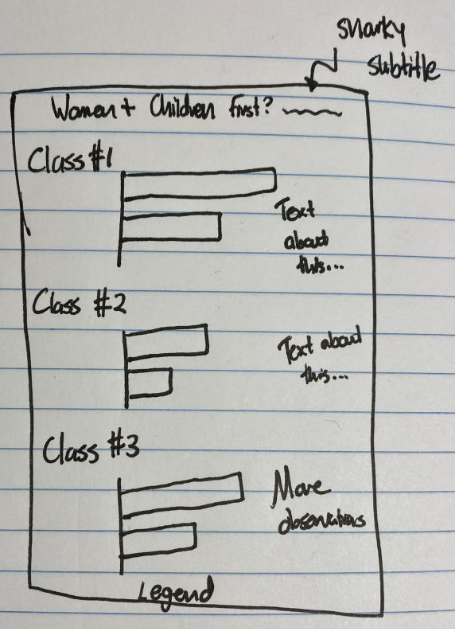
# Introduction

The Titanic disaster database from Kaggle is a dataset containing passenger information and survival outcomes for all of 2,224 passengers and crew [1] aboard the ship. It includes a total of 10 variables and 2224 observations; a complete data dictionary is available in [1]. For our analysis, we used the test dataset available from Kaggle, which gives us a substantial number of observations available for analysis.

# Good Visualizations

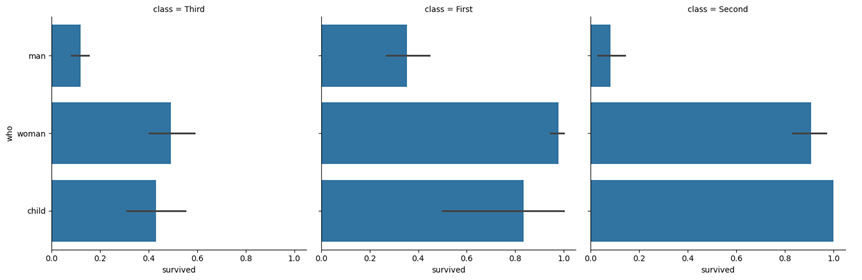
## Good Visualization #1 – Women and Children (And Rich Men) First? Analyzing Survival Across Ticket Class on the Titanic

For my graph, I wanted to focus on the different survival distributions across gender and economic class. I knew that women and children were more likely, but I wanted to explore how this survival rate varied across economic class. Since I was comparing survival rates, I knew that I would want to use a bar chart, and since I was breaking the analysis apart on passenger class, I wanted to use small multiples. In my original design diagram, I was trying to accomplish an infographic type design, with callout values calling attention to some of the insights I found in my data exploration.



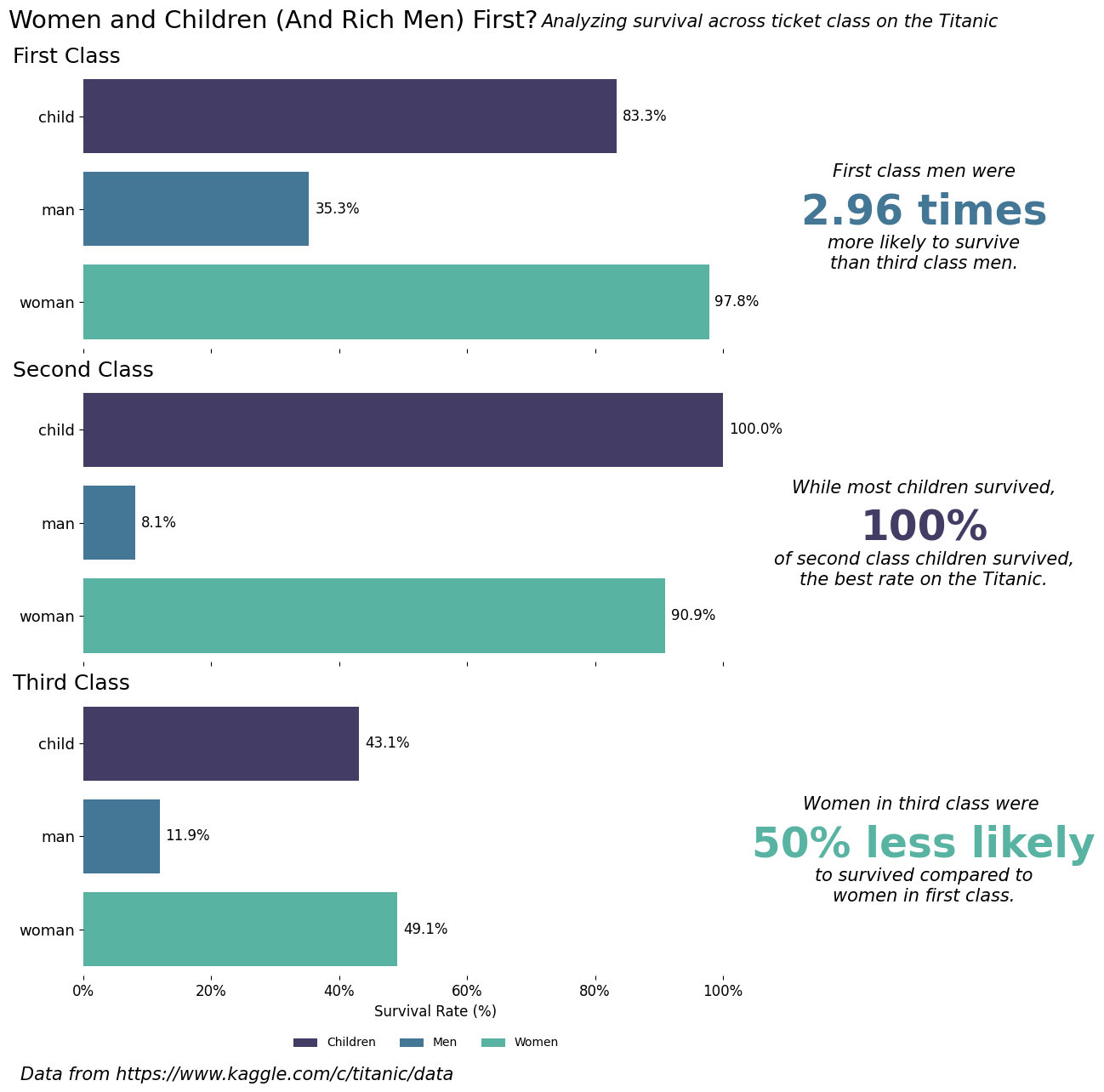
**Figure 1:** Preliminary Design Diagram for *Women and Children First?*

Unfortunately, my first attempt at creating this graph didn’t go particularly well. While it is not misleading enough to be considered one of the “bad” graphs in this report, it suffers from some problems. The horizontal layout makes it difficult to compare the relative proportions across different classes – for example, it’s not easy to tell whether first class women have a higher survival rate than second class women. The error bars are confusing and it is not easy to quickly differentiate between genders.



**Figure 2:** Initial Attempt at Creating *Women and Children First?*

After my first attempt, I created the final version of *Women and Children First?* For this visualization, I switched the chart layout to vertical instead of horizontal (which surprisingly isn’t easy to do in Matplotlib?). I feltlike stacking the bars vertically makes them easier to compare – now the reader can easily tell the difference between first and second class women’s survival rates. I also incorporated color to easily differentiate between children, men, and women, added labels to enable accurate comparison of values, and added a legend. Also, since visualizations should be as self-sufficient as possible, I added infographic callouts to emphasize important insights and included a source for the original dataset.



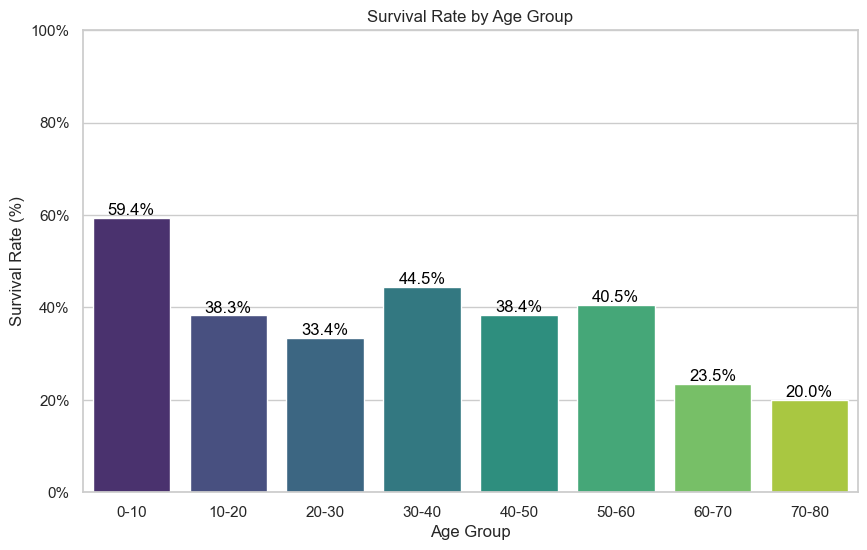
**Figure 3:** ***Women and Children (And Rich Men) First?*** *Analyzing survival across ticket class on the Titanic*

Good Visualization #2 – Survival Rate by Age Group

With the Survival Rate by Age Group visualization, I wanted to explore if there was a relationship between the age of passengers and their survival rate. I utilized a bar plot to explore this relationship between age and survival rate. I added percentages above the individual bars, so it was easier to see the actual survival rate. I used the viridis color palette for better accessibility for colorblind people.

Regarding the information in the visualization itself, we can see that people aged 0-10 had the highest survival rate. From this, we can infer that young children were likely given priority when evacuating. Interestingly enough, the elderly had significantly lower survival rates compared to the other age groups.

Overall, this graphic does an adequate job in displaying survival rates across age groups. It conveys information in an accurate and simple manner without being overly flashy or confusing. It also considers accessibility for colorblind people by using a colorblind friendly palette.



**Figure 4: Survival Rate by Age Group**

# Bad Visualizations

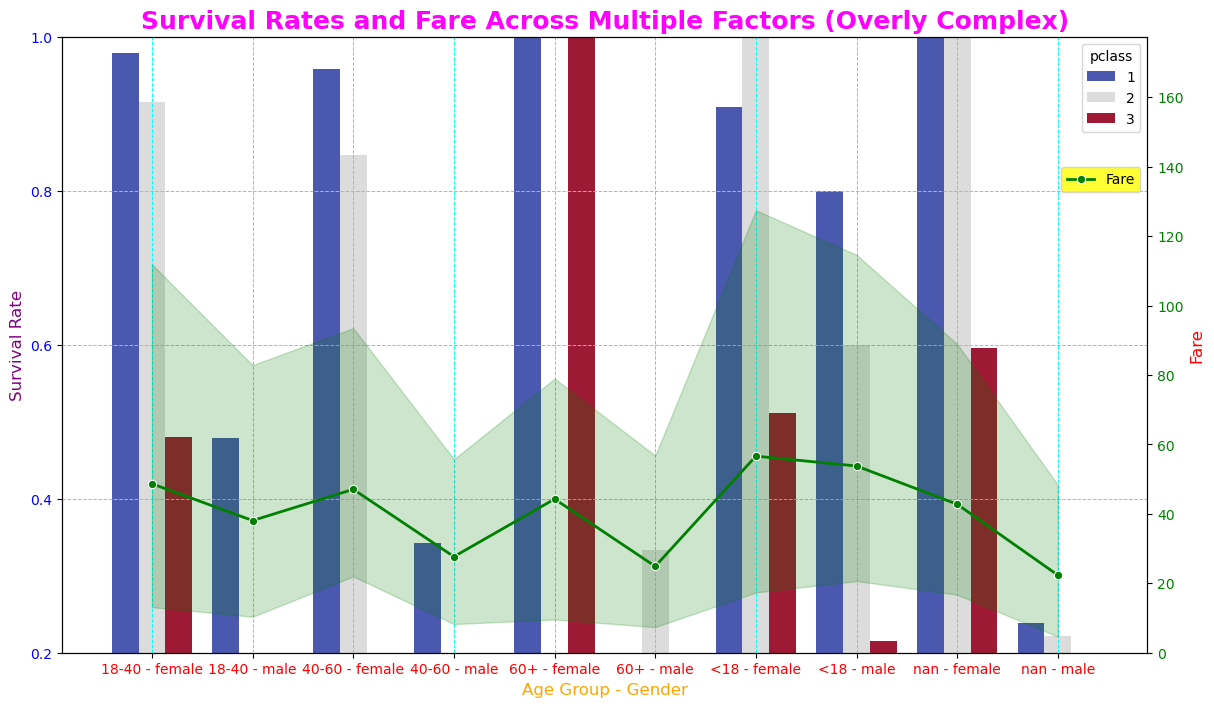
## Bad Visualization #1 – Survival Rates and Fare Across Multiple Factors (Overly Complex)

The visualization below tries to show how survival rates and fare prices change across different age and gender groups, split by passenger class. While the goal is to present a comprehensive picture of the Titanic dataset, there are some key issues with the design that makes this visualization misleading and a bit confusing.

One of the most problematic aspects of this chart is the manipulation of the y-axis range for the survival rate. By setting the y-axis to start at 0.2 instead of 0, the differences between survival rates appear far mor significant than they really are. For examples, an change in survival rate from 0.35 to 0.55 looks much more dramatic due to the compresses scale.

Another issue is the unnecessary complexity introduced by combining both bar plots for survival rate and a line plot for fare introduces an irrelevant variable, making it seem like survival rates and fare are linked when they are not. This dual-axis presentation is a distraction from the main analysis, leading to an inaccurate interpretation that survival rates and fare are more connected or significant across groups than they actually are. The complexity of the chart makes it hard to understand what’s really going on.

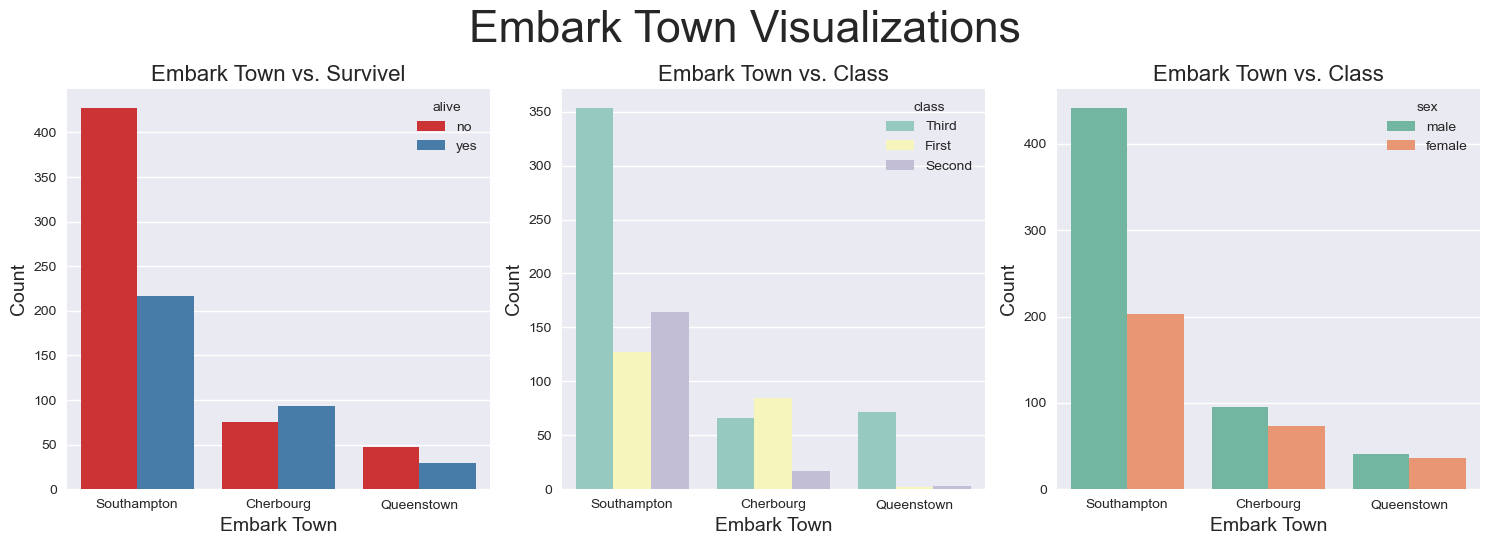
Additionally, the use of multiple colors and a cluttered layout with bright colors for the title, gridlines, and chart elements adds confusion. The cyan gridlines and magenta title are visually striking, but they draw attention away from the data and make the chart feel overly busy than what it already is. Viewers may struggle to focus on the key insights due to the overwhelming use of color.

**Figure 5**: Survival Rates and Fare Across Mutiple Factors (Overly Complex)

## Bad Visualization #2 – Embarktown Matters for Survival Rates – Must be Good Swimmers!

The "Embark Town vs. Survival" visualization might initially suggest that the embarkation town played a critical role in determining a passenger’s fate. Looking at the first chart, one could easily conclude that people boarding in Cherbourg had a significantly higher survival rate compared to those from Southampton. However, this is a classic case of correlation not equaling causation. The real factor at play isn’t the embarkation point itself, but the class distribution of passengers from each town—something that isn’t explicitly shown here.

Cherbourg happened to have a higher proportion of first-class passengers, who were more likely to survive due to better access to lifeboats, while Southampton had a higher share of third-class passengers, who, unfortunately, faced a much tougher escape. So, unless you believe that Cherbourg passengers were secretly trained in maritime survival, this chart is misleading. And while we can’t rule out the possibility that they were simply better swimmers, it’s probably more accurate to credit their extra cash for getting them a prime spot on the lifeboats!



# Ethical Concerns in Visualizations

Data scientists have the ethical responsibility to make sure that their visualizations are clear and easily interpretable, making sure that they can be easily understood by anyone, whether it’s ensuring accessibility or making sure the visual design is simple enough to not confuse viewers.

One of the main concerns focused on while creating *Women and Children (And Rich Men) First?* was color-blind accessibility. Ethically, data scientists need to ensure that their visualizations can be understood by anybody, regardless of their genetic makeup or education levels. As a data scientist, you should not knowingly create visualizations that could mislead a significant proportion of the population / readers. Throughout the development process of *Women and Children*, both software tools for testing color-blind accessibility and the more informal test of asking color-blind family members if they could differentiate the colors used in the visualization were used to help ensure colorblind friendliness [2].

Additionally, the chart in figure 5 shows how visual design can be manipulated to mislead viewers. By distorting the y-axis, overloading the chart with multiple irrelevant data points, and using distracting colors, the visualization fails to accurately convey the data. For this graph, a more ethical approach would involve simplifying that chart by using a consistent scale for the survival rate and avoiding unnecessary complexity that could confuse the viewer.

# References

[1] “Titanic - Machine Learning from Disaster.” Accessed: Feb. 24, 2025. [Online]. Available: <https://kaggle.com/titanic>

[2] “jsColorblindSimulator.” Accessed: Feb. 24, 2025. [Online]. Available: https://mapeper.github.io/jsColorblindSimulator/