The Titanic: “Women and Children First”

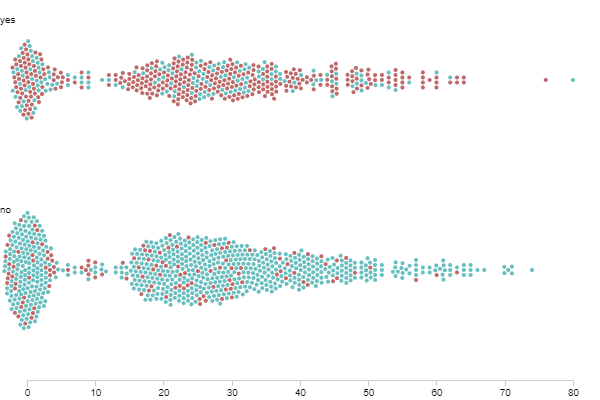
Jennifer Bochenek, Toby Obunadike, Joan Kibaara

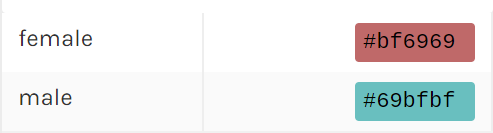
Spring 2020

**Introduction**

Using the sample dataset in RawGraphs for Titanic Passengers, we chose to review if passengers actually followed the oft-touted mandate of “women and children first” when it came to loading the lifeboats. We chose two different graphs to look at this issue, a Beeswarm plot and a Sunburst chart, then provide a brief analysis of the visualization and what we learned from each approach.

Figure 1. Beeswarm Plot of Age by Survival and Gender





*Caption: Age as the x-axis, where unknown age shows up at 0, and then goes from age 1 all the way to 80. The two groups show whether or not they survived the sinking of the Titanic. Within each group, there are two different colors, which are for the two genders in the data set. In this graph, the burgundy represents women, and teal is for men.*

**Process of building the Beeswarm plot**

To build the Beeswarm plot, we used the variables age, survival, and gender. Age was used as the x-axis value, survival as the groups, and gender as the colors, which resulted in figure 1.

**Analysis of the Beeswarm plot**

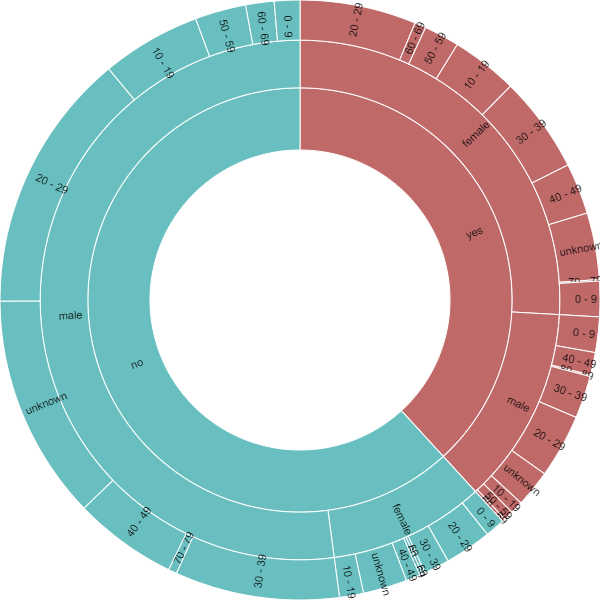
If the passengers of the Titanic followed the “women and children first” mandate, we would expect to see more females (more burgundy dots) in the “yes” group than in the “no” group. This holds with the data shown, as the “yes” graph is mostly red with smaller amounts of blue, while the “no” graph is mostly teal with very little burgundy. It is important to note that men make up 64% of the entire passenger list, so if men and women survived at an equal rate, then we should expect to see both graphs being slightly more blue than red. For determining whether or not more children survived, one has to compare the ‘thickness’ of the survived (yes group) and died (no group) to one another across the age groups on the x-axis. In Beeswarm plots, data is shown in a non-overlapping manner, which is particularly helpful when there can be many data points in the same place. It overcomes the issue of overlapping points by showing points alongside each other, so the thickness or width shows the distribution or number of points in an area. Thus the shape of a Beeswarm plot becomes part of the interpretation. In this case, the thickness of the surviving group is larger for under approximate 15 years of age, with a few exceptions. For example, it appears that more died around the age of 10 then survived. One problem with comparing the two graphs is that more passengers died (62%) than survived (38%), so as a result, the dead group will naturally be thicker. However, this also means that if the surviving group is thicker, it is particularly notable. Thus we can also say that in general more children survived than died.

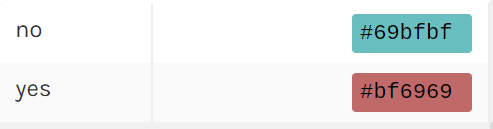
Thus, we can say that they did seem to follow the general concept of “women and children first” as the survival category holds far more women than men even though men outnumber women passengers, and the bulk of the children (0 being unknown age and not a child) are in the survival category.

**What we learned from the Beeswarm plot**

Another key takeaway is that a Beeswarm plot has some difficulty in interpreting differences between two swarms when there is an imbalance in sizes of the overall swarms. As well, it can be difficult to determine a proportion within each swarm, as the points are arranged along the x-axis and not grouped up, thus because they are dispersed throughout the swarm it is hard to make an accurate comparison between the male and female groups. One benefit of the Beeswarm is that the shape itself allows interpretability, but that does not outweigh the negatives for the purposes of this analysis. To overcome these issues, we also used a Sunburst chart to examine the same data.

Figure 2. Sunburst Chart of Survival by Gender and Age Group





*Caption: The chart uses three variables, survival, then gender, then age group in that order from the inner ring to the outer ring. The color in the graph represents those who survived as burgundy and those who did not as teal.*

**Building the Sunburst chart**

To build the Sunburst chart, we used the variables age group, survival, and gender. Survival was placed first in the hierarchy followed by gender and last age group. Survival was also added to the color attribute, which resulted in figure 2.

**Analysis of the Sunburst chart**

The innermost circle can visually tell us the survival rate of the passengers. As we can see, there are more passengers who did not survive the sinking of the Titanic than who did. Those that survived are approximately ⅔ of the total.

Next, it is possible to determine the ratio of male to female passengers and their survival rate. This can be determined from the middle circle. From the chart, we can see that there were more males who did not survive. The teal color has a larger section labeled “male” compared to the section labeled “female”. The burgundy color has a larger section labeled “female” compared to “male”, even though overall males formed a much greater portion of the passengers.

Last but not least, we did an analysis of the passenger age groups and their survival. This can be determined by the outermost circle in the chart. Generally speaking, the majority of the male who did not survive the crash fall in the 20-29 and 30-39 and 40-49 age groups. We can also see that the youngest passengers who died in the crash were all-male in the age group 0-9. The more elderly passengers (age groups 70 -79) died in the crash, while the few who survived were all female.

Similarly, most of the surviving passengers fall in the 20-29, 30-39 and 40-49 age groups. The youngest surviving passengers fall in the 0-9 age group and are both male and female, while the oldest surviving passengers fall in the 80-89 age group.

**What we learned from the Sunburst chart**

The Sunburst chart gives us a better visual of the different breakdowns and we can easily determine the proportions. For example, we can see that about ⅓ of the passengers survived the crash. One disadvantage we noted about this chart is that when you have many levels in a variable, it becomes difficult to fit the variable in the chart. In our example, we chose to use the “Age Group” variable as opposed to the “Age” variable, since the “Age” variable had much more levels.

**Reflection on what we learned**

The Beeswarm was useful in showing where the bulk of the data was because the shape of the beeswarm was an interpretable feature, while the Sunburst was useful in clearly showing proportions of the data. Thus each of them has a strength that makes them useful for this analysis, and each had disadvantages that would make them difficult to interpret.