Gourav Aich

Data Visualization: Studying Titanic Survival

Summary

This project is about creating a data visualization using Tableau that tells the story of Titanic survivors. It depicts the number of survivors and their survival chances based on certain characteristics like age, sex and class. By using various graphs, it draws conclusion that the men onboard had the lowest survival rate when compared to women and children. However, class seems to be a far weaker variable in determining survival rate than sex or age.

Design

As suggested by Udacity, I downloaded the train dataset from Kaggle at https://www.kaggle.com/c/titanic/data, and found that except some missing values for "*Age*" and "*Embarked*" columns, the data was clean and wouldn't require much data wrangling.

After exploring the data set, I decided to perform a couple of transformations that would help me in crafting a message and a related story around my data. Among these transformations, *I would particularly like to explain how I categorised the passengers as Children based on Age, Sex and Name columns*.

Categorising passengers as Men, Women and Children - For all those male passengers with missing Age values in the dataset, I categorised the ones having their title as "Master" as children. However, for female passengers with missing Age, there wasn't any such identifier. So, I categorised the ones having "Sib Sp" column value of at least 1 and "Parch" column value not equal to 0 as children. The idea was to identify those female passengers travelling with siblings and not spouses (hence, Sib Sp > 1) AND travelling with their parents (hence, Parch not 0). It's worth mentioning here that there are some children with Parch=0 who travelled only with a nanny. However, there is no way I can differentiate such records from the ones who were adults and travelling alone.
Sib Sp - # of siblings / spouses aboard the Titanic
Parch - # of parents / children aboard the Titanic

Now, coming to the passengers with Age, I manually scanned the dataset and found that males till the maximum age of 13 were addressed as Master. So, I categorised all passengers whose age was 13 and below as children.

Embarkation by Class - Once the data was transformed and ready for use, I thought about narrating the story right from the time when the passengers boarded the Titanic. I considered showing the number of passengers by port of embarkation using a bubble chart as well as number of passengers by class and port of embarkation using side-by-side bars. That way, not only was it clear that most of the passengers boarded from Southampton, it also showed the distribution of these passengers class-wise.

Effect of gender and class on survival - Next, I wanted to show the effect of gender-category (i.e. Men, Women and Children) and class on survival. To represent the survived/dead number of passengers across each gender category as well as across each class, I used a bar chart for each of them on a single dashboard as it provides better insights while comparing various entities. To depict the survival status, I used colour whereas I added the class and gender-category as filters. The idea was to provide the option of taking a closer look at the data depicted by the visualization by mixing and matching gender and class values.

Effect of age on survival – The next step was to explain whether age played a significant role in survival chances. Age being a continuous data type ranging from 0 to 80, a histogram becomes the natural choice to represent the age distribution. So, I converted age into discrete bins of size "2" and marked the survival status by colour which will help in knowing if at all passengers belonging to certain age groups had a higher chance of survival or vice-versa. I also

provided the gender-category as a filter so that one can dig deep and take a closer look by each of the men/women/children category and find out if age affected the survival of any/some/all of these categories.

Detailed data-view of each passenger – For all the 891 passengers of this data set, I wanted to show the survival/dead count for each category by class attribute. A tabular representation including a column grand total seemed to be the right fit to display this information. Additionally, I decided to show the survival status, name and age of each of the passengers. For this, I marked the survival status using colour and provided it as a filter, too, whereas for category (i.e. men/women/children), I used shapes. The objective was to provide a granular view for each of the passengers as well as a holistic view by category and class.

Initial version shared for feedback:
https://public.tableau.com/profile/gourav1439#!/vizhome/TitanicViz_2/Story_Initial

Feedback

I shared my visualization with a colleague of mine to get feedback. The screenshot of the email conversation is provided below.

From Deepti Mishra Hide

To Gourav Aich

Hi Gourav,

Thanks for sharing your Titanic story. I took a look and this is what I think:

1. As this story is about the Titanic survival, the visualization about embarkation point seems to be not so relevant. I feel it kinds of distracts the attention away from the main study.
2. Also, I think the 2 graphs regarding "effect of gender and class on survival" on the same page kept on confusing me. Won't it be better if you keep them in separate slides and provide some additional info to give a clearer picture of their effect? What do you think??
3. In the age-related graph, there are more than 200 passengers whose age is "blank". That is almost 25% of the total passengers. Not sure I understand what this graph is trying to convey.
4. The last slide of the story is the one that I found to be most interesting. Probably providing the "class" detail about each of the passengers when I hover on each of them will be good, though.

One more thing that I want to point out about this story is the "title" on each of the slides. There isn't much continuity between two consecutive titles. Because of which, it doesn't seem to be telling a story.

Regards,
Deepti.

From: Gourav Aich
Sent: 26 November 2017 01:15
To: deepti.mishra@hotmail.com
Subject: Please provide feedback

Hi Deepti,

As we had earlier discussed, can you please take a look at my Tableau story on Titanic Survival Analysis and provide feedback.

Here is the link:
https://public.tableau.com/profile/gourav1439#!/vizhome/TitanicViz_2/Story_Initial

Thanks,
Gourav.

Gourav Aich

Final Visualization

Post-feedback, I implemented the suggestions and made some additional changes to arrive at the final visualization.

Firstly, I removed the embarkation visualization so that my story stay focused on survival. Additionally, I removed the age-related graph as there were too many passengers with *blank* age and I didn't want to add "junk". Moreover, the categorisation of passengers should logically take care of this aspect.

I also un-cluttered the story by splitting the category (*men/women/children*) based visualization and class-based visualization, and providing filters for each of them. A small table was added to each of them showing the percentage of passengers who survived or died. Applying category/class filters would alter the percentage accordingly and will show how good/bad the survival chances were for the filtered values.

It was now clear that category played the most important role in a way that the "men" on board had the least chances of survival. I felt I should take a closer look and therefore, created another plot to find out the percent of passengers saved for each category. The overall survival rate for men was 16% whereas for women, it was 77%, and for children, 54%. Clearly, it was "women and children first". *Additionally, it corroborates the incident about the captain of Titanic threatening to shoot men who got into the lifeboats* (https://goo.gl/bxN8Vr).

Class, on the other hand seemed to have lesser effect on the chances of survival. To get a deeper insight, I combined category & class and plotted the percentage of total passengers saved for each combination. This way, I came to know that men from second class had the worst survival rate of 8%, followed by men from third class (11%) and first class (35%) respectively. Though the previous plot showed that women, overall, had a better survival rate than children, this one revealed that the children of second class had the best survival rate of 100%.

Lastly, in the passenger's "detailed data-view", I added class and gender to the tooltip information so that the children's gender can also be viewed. To give a proper narrative to my tableau story, I updated the titles in a way that they act as headlines to the visualization that follows.

Final version for submission:
https://public.tableau.com/profile/gourav1439#!/vizhome/TitanicViz_2/Story_Final

Resources

https://www.livescience.com/19635-titanic-disaster-happen.html
https://en.wikipedia.org/wiki/RMS_Titanic
Udacity's "*Data Story Telling*" course