Data Visualization

Kristen Forte, Tanviben Goswami, Keerthana Lingala, Sebastian Segura

Introduction

Data Visualization is a critical skill to master to gain the ability to shape how audiences understand information. Throughout the project, the visualization will be analyzed, and good vs bad will be compared. Good Visualization can provide audiences with a clear insight into the information being conveyed, so that viewers can have a clear understanding and draw better conclusions, especially when they are not a highly technical audience. Bad Visualizations will do the complete opposite, which will just confuse the audience and lead to a wrong understanding of the information, resulting in incorrect conclusions.

This project will show a couple of examples of good and bad visualizations from the Titanic dataset that follow the best and bad practices that effectively communicate data information, and where the bad visualizations are purposely created to highlight all the common errors and confuse audiences. The Titanic dataset contains demographic data and information about passenger survivors, which will be used for the analysis of survivor patterns across aspects like gender and age. Using the dataset to highlight and show multiple differences between good and bad visualizations, and explain the ethical implications of misleading data, and how this can affect the truth of the information being presented.

Data Information

The data used for this project is the Titanic dataset from Kaggle (https://www.kaggle.com/competitions/titanic/data). The dataset consists of information about the passengers of the famously doomed 1912 voyage of the Titanic. The various information includes:

- survival
  - 0 = No
  - 1 = Yes
- pclass
  - 1 = 1st
  - 2 = 2nd
  - 3 = 3rd
- sex
- Age
  - In years
- sibsp
  - Number of siblings/spouses aboard
- parch
  - Number of parents/children aboard
- ticket
  - Ticket number

- fare
    - Price of ticket
- cabin
    - Cabin number
- embarked
    - C = Cherbourg
    - Q = Queenstown
    - S = Southampton

Good Visualizations

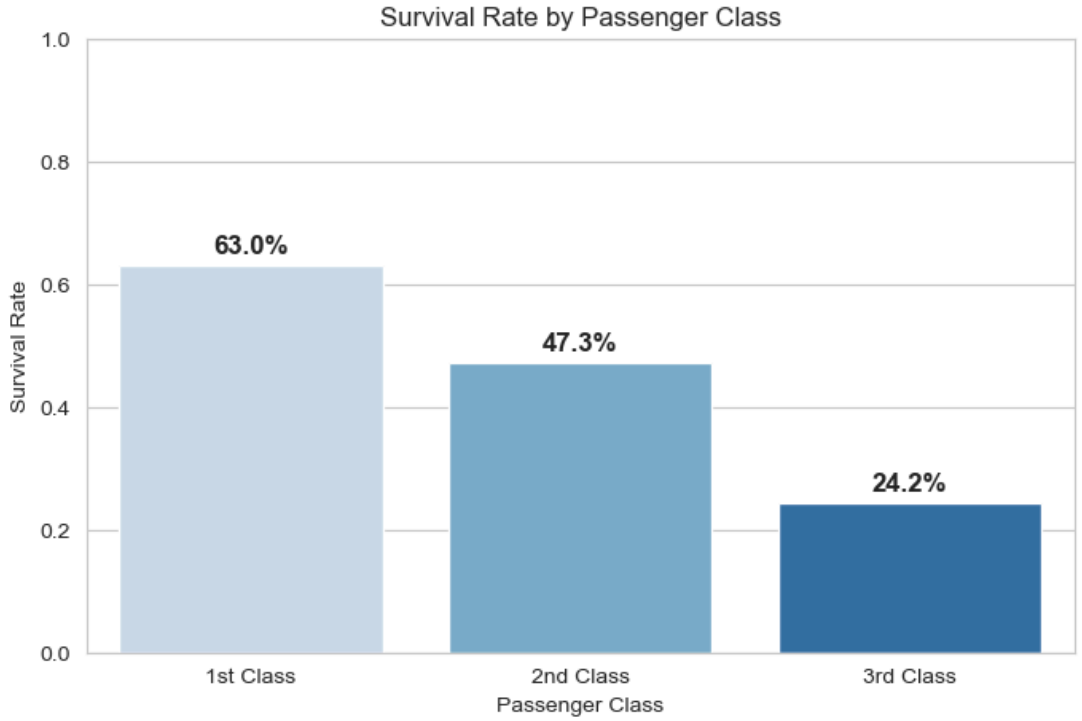

Figure 1: Survival Rate by Passenger Class. This figure clearly conveys the message represented by this bar chart. Showing the survival rate by passenger class in a descending order so that all types of audiences can interpret. Why is it effective? The design allows for easy interpretation and shows clear labels that compare the survival rates.

Figure 2: Overall Survival Proportion. This pie chart illustrates the total percentage of the passengers who survived compared to those who did not survive. The pie chart shows that 38.4% of the passengers survived, while 61.6% did not survive. The chart is labeled correctly, and the percentages are marked on the chart, which makes it simple to see the total result of the disaster. The chart does not misrepresent the percentages or conceal the categories, and it gives a clear representation of the total survival rate.

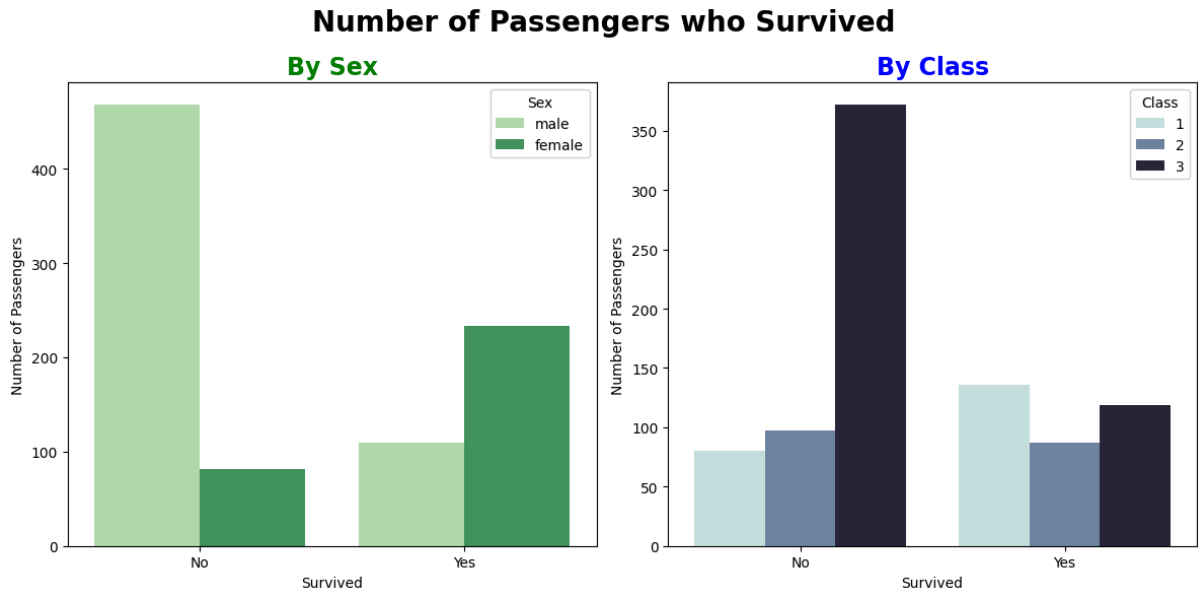

Figure 3: Number of Passengers who Survived. These bar graphs clearly display the relationship between a passenger's sex/ticket class and their survival. Titles, x and y axis labels, and legends all contribute to the interpretability of this graph.

Bad Visualizations

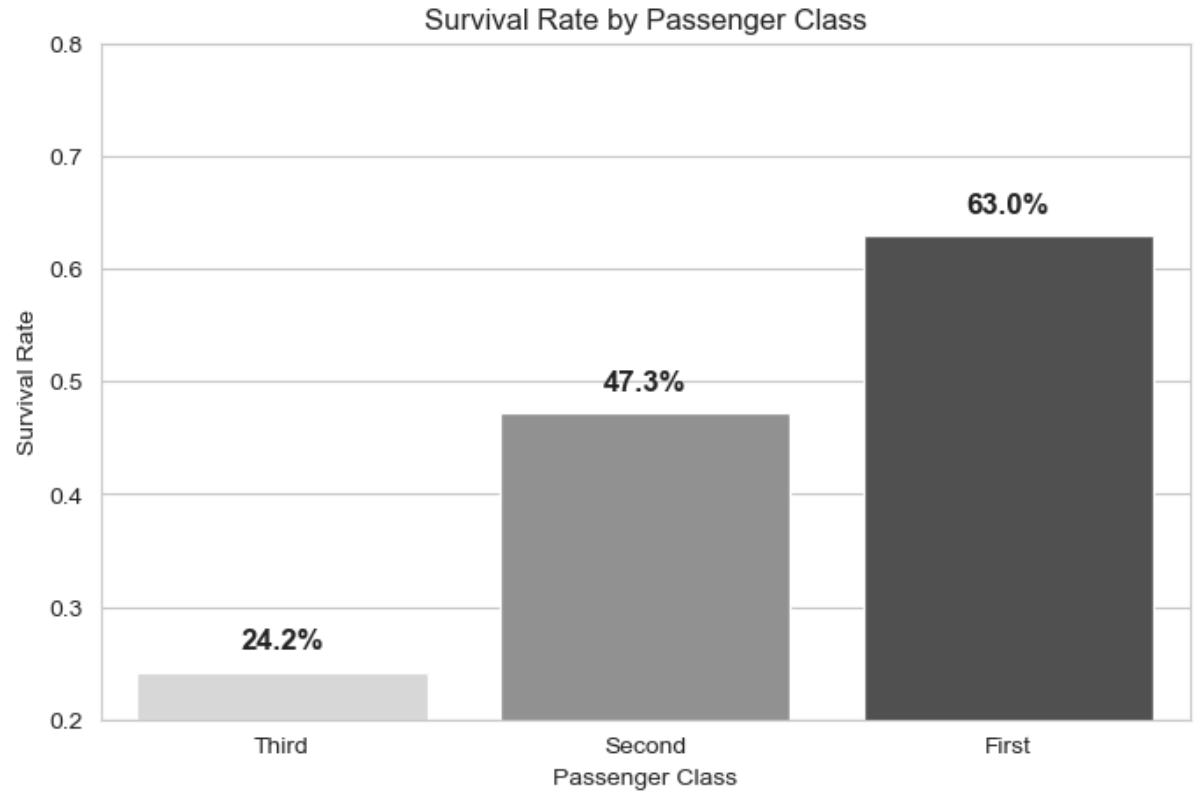

Figure 1: Survival Rate by Passenger Class. What makes this visualization not effective is that passenger classes have been altered to confuse and mislead the message. The y-axis was truncated, making the information appear more significant since it cuts off parts of the scale, which makes the

percentages in the visualization look more drastic than they actually are. Finally, the scale of tones of grey was chosen to make the viewer confused when it comes to the overall results and distinction.

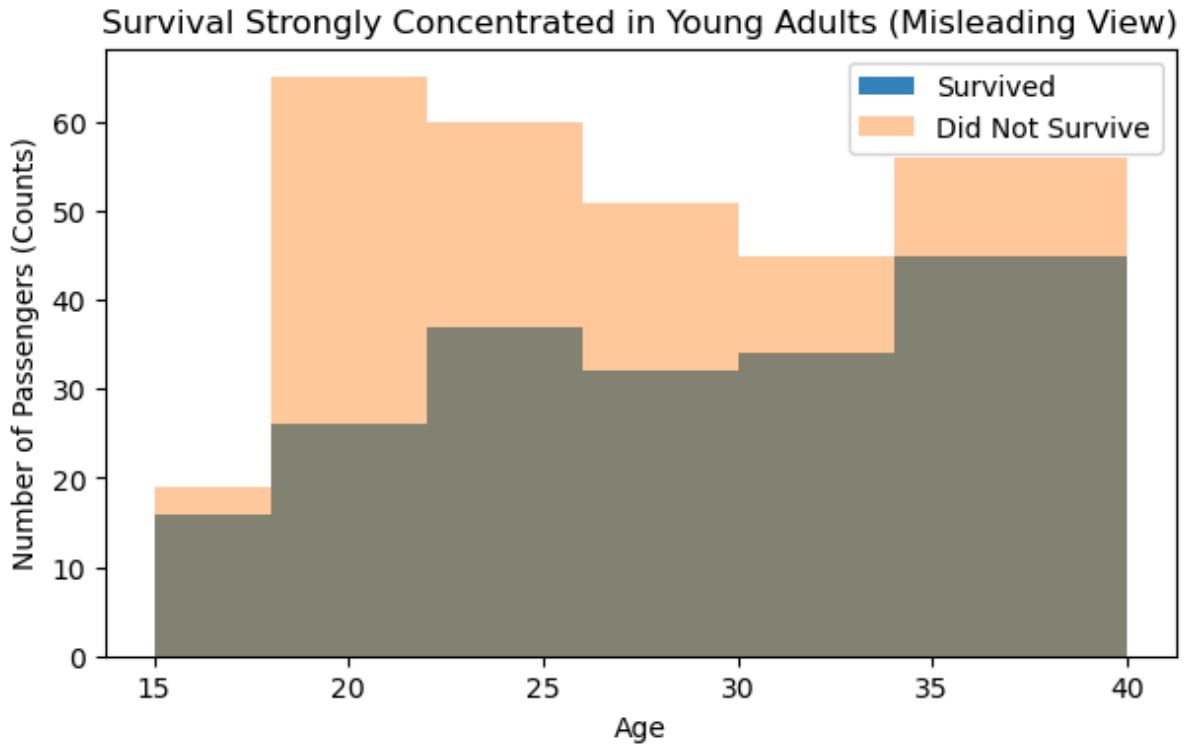

Figure 2: Survival Strongly Concentrated in Young Adults. This histogram is misleading because it only includes passengers between the ages of 15 and 40. By removing children and older adults, the chart hides important information. Since children had a higher survival priority on the Titanic, leaving them out changes the full story. It also shows counts instead of survival rates, which can confuse viewers about what is actually being compared. The way age groups are divided can also change how the patterns look. This example shows how selecting only certain data can create a biased message. Even if the data is real, leaving out important groups makes the visualization incomplete and not fully honest.

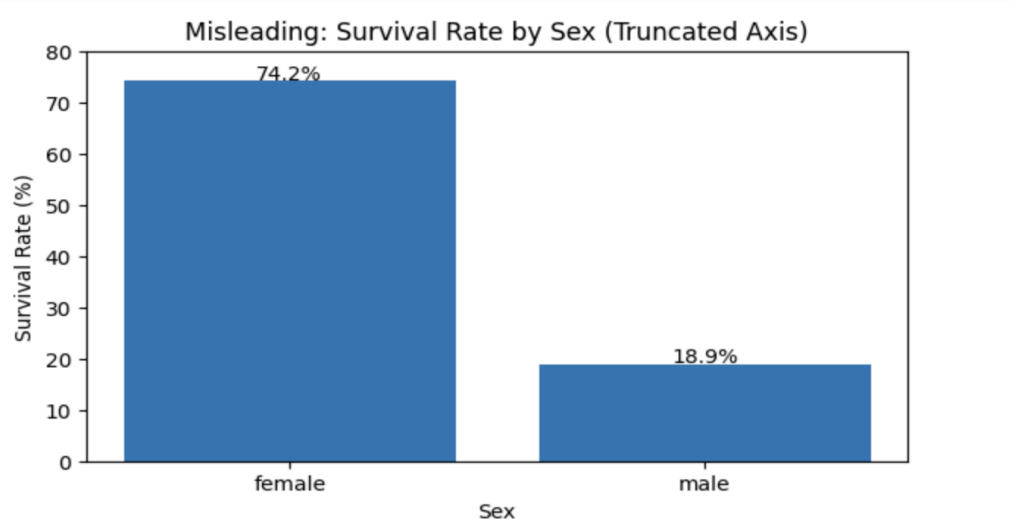

Figure 3: Survival Rate by Sex. Though this graph is employing accurate survival rate information based on gender, the y-axis has been deliberately shortened, beginning at 60% rather than 0%. Such a manipulation of the axis scale tends to emphasize the disparity between the survival rates of both genders. In this case, the survival rate for females is made to seem much greater compared to the actual disparity when a proportional perspective is taken. Even if the information presented is accurate, the axis manipulation can tend to deceive the audience into thinking that the disparity is greater.

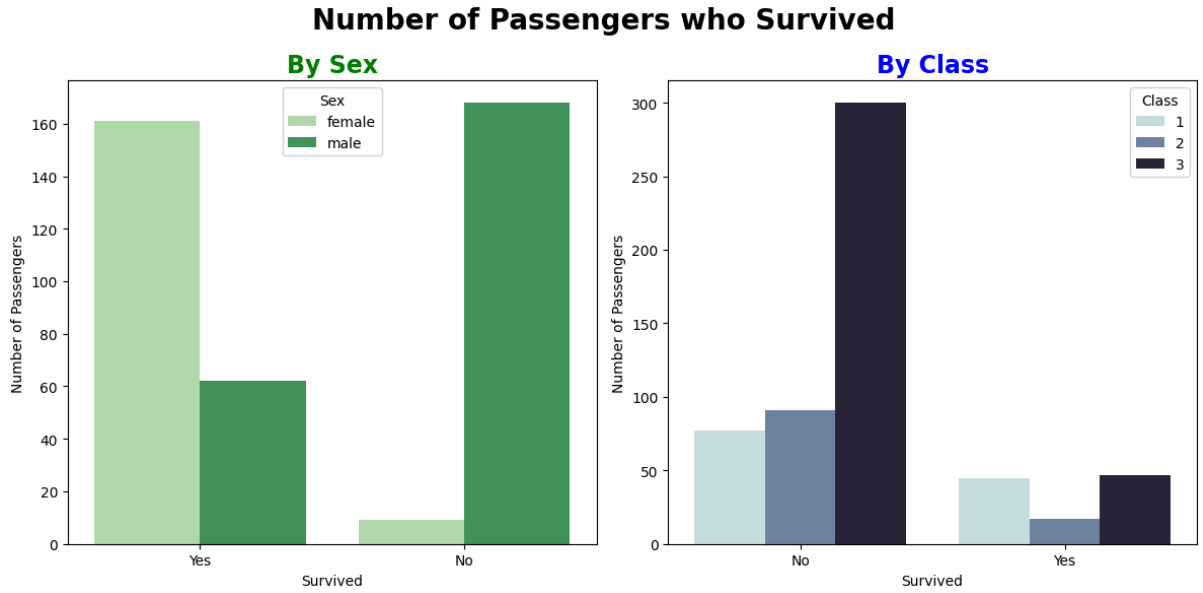

Figure 4: Number of Passengers who Survived. These are very dangerous graphs. On the surface, they look perfectly normal, even accurate, with good data representation. However, the datasets used to create the graphs have been altered. In the first graph, all 3rd Class passengers have been removed, meaning this graph only represents counts of 1st or 2nd class passengers. In the second graph, all female passengers were removed, which causes the graph to show a different (lower) survival rate.

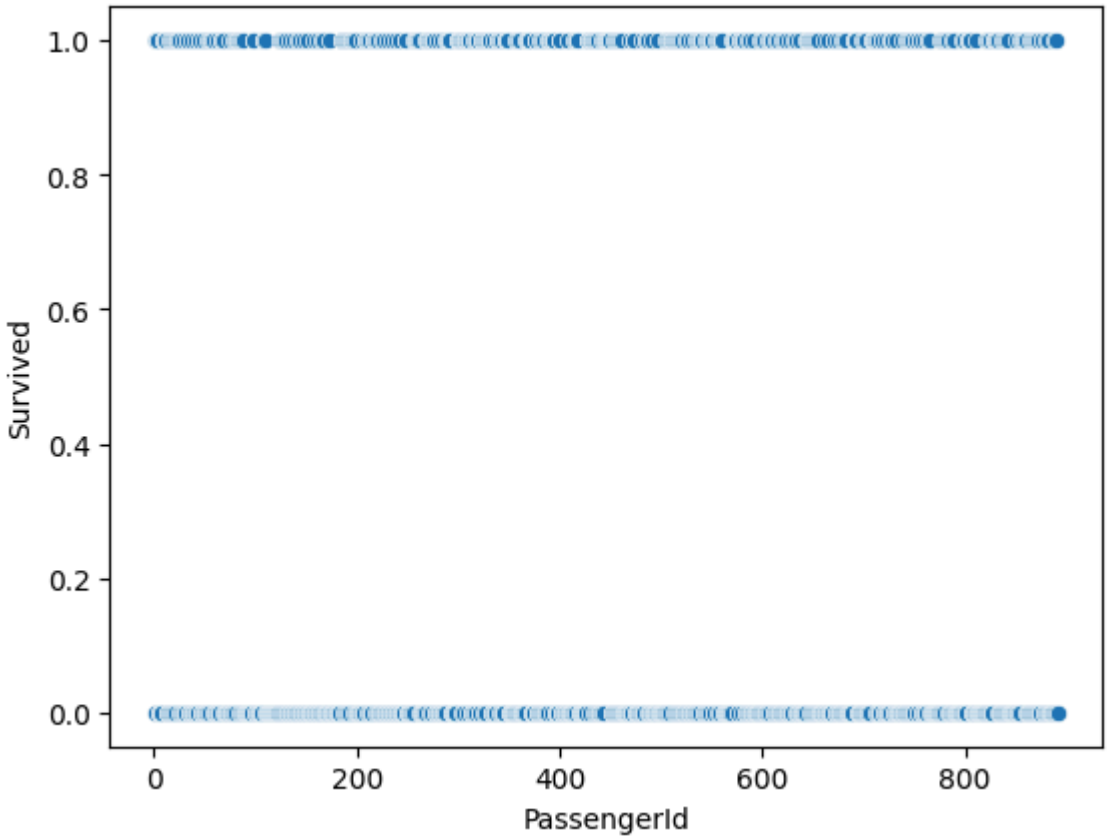

Figure 5. This graph is very noticeably bad. It does not contain a title, and what little information it does offer is very hard to interpret. The dots are so clustered together, it is basically impossible to differentiate between them.

Conclusion

The differences between what makes a visualization good versus bad can be subtle, yet very powerful. A good visualization requires effort, in both the preparation and execution of the graph. A bad visualization, however, is generally simpler to make. Care is not taken with either the type of graph used, or the data that is displayed. Some graphs are very easy to further manipulate, which leads to biased (or outright incorrect) results. These graphs can be easily used to unduly influence the viewers in many different ways.

While it may not be easy, care must always be taken to make visualizations precise, clear, and honest. Neglecting to do so shows a clear compromise in a researcher's data ethics.