DATA 6550: Visualization Project

Group 6

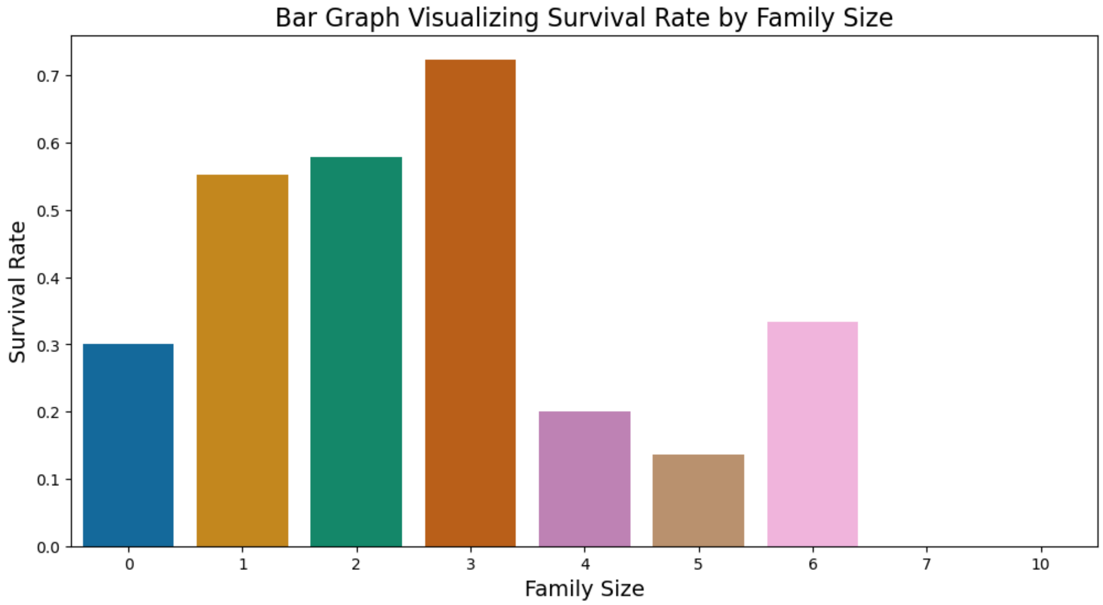
Visualization Ethics and Communication

Introduction

This project explores the impact of data visualization by analyzing different dataset variables. Each group member focused on a specific column vs. survival rates, creating both effective and misleading visualizations to examine how design choices influence data interpretation. The goal was to understand the importance of clear, accurate visualizations and the risk of misinterpretation. Poorly designed charts can distort information, leading to confusion or false conclusions, while well-structured visuals enhance clarity and decision-making. By comparing good and misleading examples, this analysis emphasizes the ethical responsibility of presenting data truthfully and effectively. For this assignment, we were tasked to create compelling visualizations of varying degrees of effectiveness and accuracy using the Titanic dataset. Specifically, we opted to use the Titanic dataset that is built into the Seaborn package using the load_dataset function. The dataset contains 891 rows across 15 columns and contains information related to the passengers of the Titanic such as: age, embark_town, cabin, passenger class, and whether they survived. Our analysis will be divided into two key segments, Accurate Visualizations and Misleading Visualizations, per the assignment guide.

Accurate Visualizations:

Autumn Noblit (Family Size vs. Survival Rate)



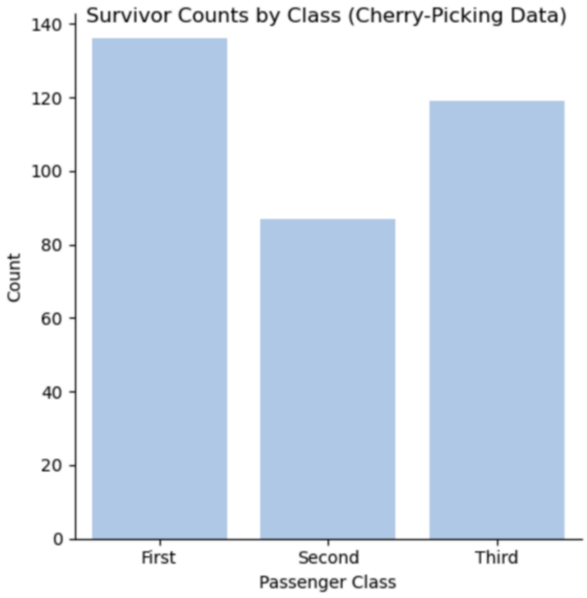
Why this is effective:

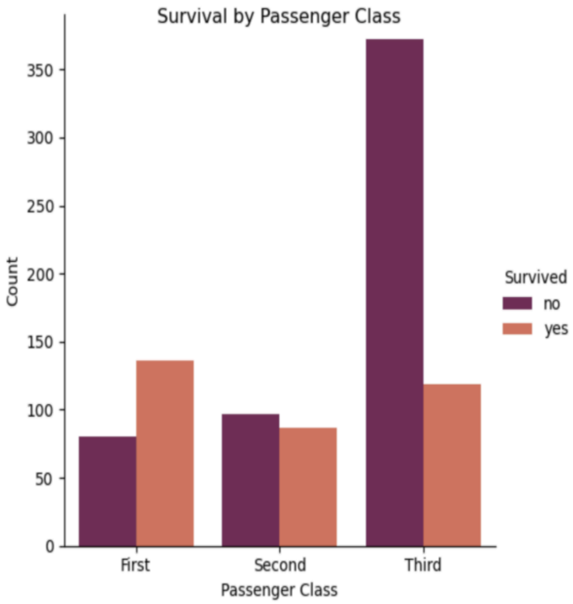
- Correct type of chart for
    - One can easily see how the various survival rates compared to families of different sizes
- Uses rate to for comparison vs. count
    - Count may be misleading as individual group counts varies
- Color palette makes it easy to distinguish groups from one another
- The y-axis begins at 0
- Chart is clearly labeled on the x and y axes and there is a chart title

By using the components of visualization appropriately, one can see that there were distinguishable differences between survival rates for families of varying size without having to spend an excessive amount of time trying to interpret the the layout of the graph.

Kirstin Tretter - Survival Rate & Passenger Class

Focus: This portion of the project focuses on visualizing the relationship between passenger class and survival and presents a comparison of two ways of visualizing this information.

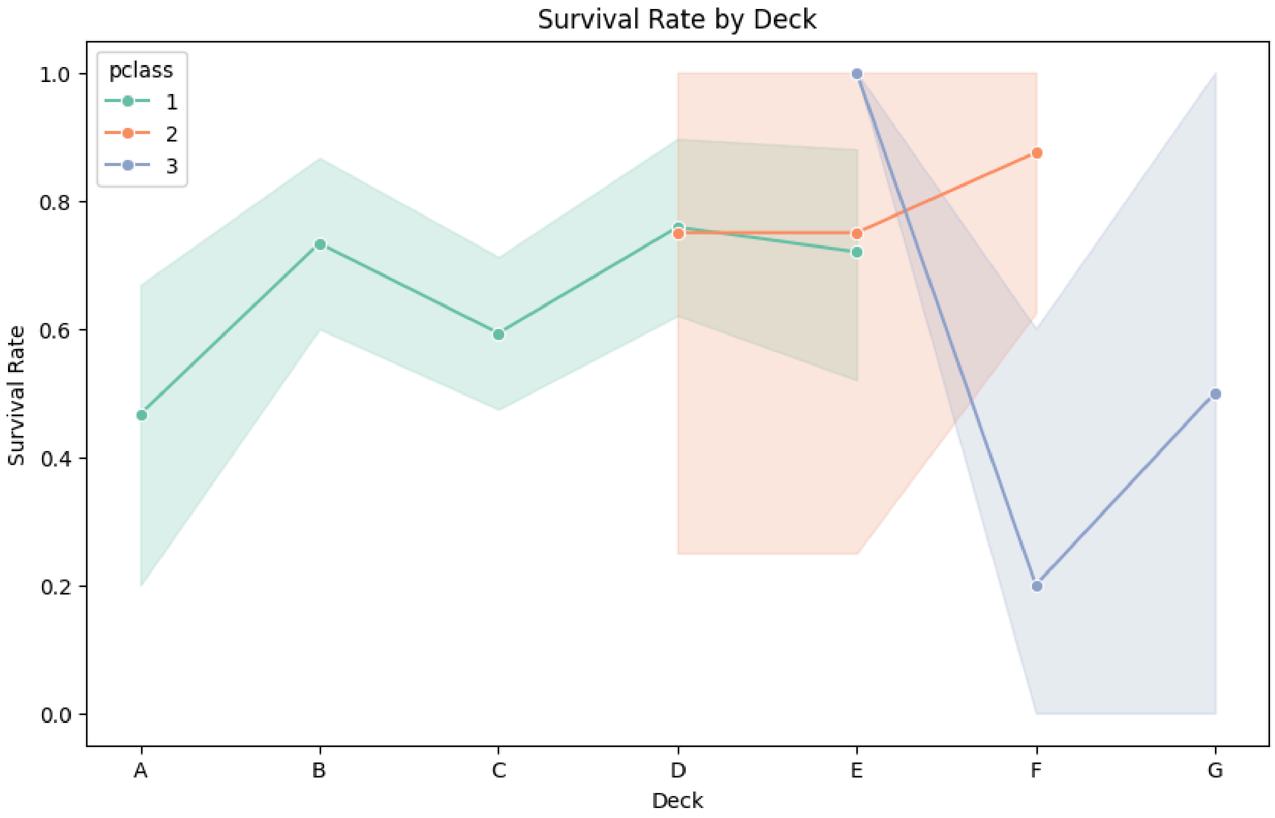




The figure on the RIGHT is an accurate portrayal of survival by passenger class and displays a count of passengers in each class that lived and died. It is an effective representation of the data including all data points, presenting a clear picture of the differences in survival between passengers of different classes. It uses an appropriate chart style for the data with color differences and a legend that enhances understanding. The reader can clearly see the apparent class disparity between the deaths of third class passengers versus first and second class. The figure on the LEFT, on the other hand, shows only the number of passengers in each class that survived the titanic disaster. This is misleading because it excludes passengers that died - cherry-picking the data. This visualization could lead a reader to conclude that there were minimal differences in survivorship between the passenger classes, or even that third class passengers had a greater chance of survival than second class. By excluding the deaths, it is impossible for the viewer to accurately get a sense of what *proportion* of each passenger class survived. The class disparity in the number of **deaths** is essential for the viewer to fully understand the disaster: 372 third class passengers, 90 second class passengers, and 80 first class passengers perished on the Titanic. Comparing the two graphs, it is clear how important including as much data as possible is for reader understanding and an ethical portrayal of information.

Misleading Visualizations:

Survival Rate by Deck (Chance):



Passenger Class	Survival Rate (Cherry Picked)	Survival Rate (Full dataset)
1	66%	63%
2	81%	47%
3	50%	24%

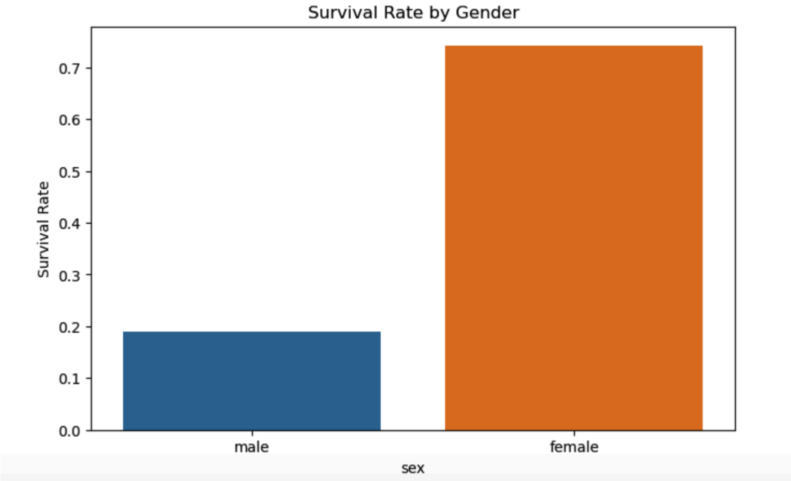
Passenger Class	Passenger Count (Cherry Picked)	Passenger Count (Full dataset)
1	173	214
2	16	184
3	12	491

This lineplot deliberately misrepresents the data by examining the 'Deck' feature to illustrate survival by Passenger Class. Upon importing the data from the seaborn, the 'deck' feature is missing approximately 77% of values; only 23% of the data is being represented. From the 23% of values present in 'deck', 173 passengers are first class, followed by 16 in second class, and 12 in third class. When compared to the full dataset, first class is overrepresented while second and third class are severely underrepresented; In total first class had 214 passengers, while second and third class had 184 and 491 passengers, respectively. Using 'deck' as a key visualization feature in this way distorts the reality of the dataset, skewing the observed survival rate of passengers by passenger class. Based on the full data set, the survival rate of passengers in first class is 63%, second class is 47%, and third class is 24%. The cherry picked data illustrates survival rates of 66%, 81%, and 50% for first, second, and third class respectively. From a design perspective, the plot leverages color, and axis labeling, appropriately to differentiate between passenger classes, but the plot itself is not a good fit for the data. Since the 'deck' feature is not continuous in nature, a lineplot is not an appropriate visualization method; a barplot or countplot would be more appropriate. From this visualization, a viewer could easily conclude that the survival rate of passengers on the Titanic was much higher across all passenger classes than it actually was. Additionally, the associated plot could lead a viewer to assume that there is a quantitative relationship between the survival rates of each cabin.

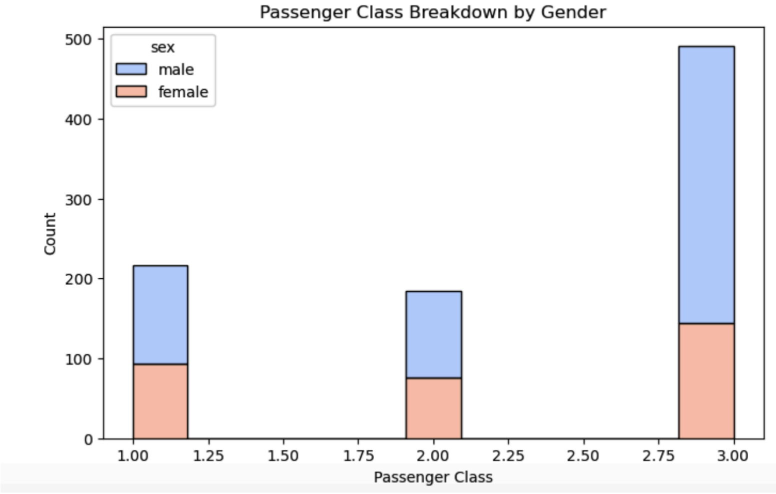
Ashlyn Brown (Gender vs. Survival)

- **Focus:** My portion of the analysis focused on exploring the relationship between gender and survival rates, while also considering other features like passenger class and age.

- **Effective Visualizations:**
  - **Bar Chart:** Showed clear survival rates differences between males and females (males 20%, females 70% survival)

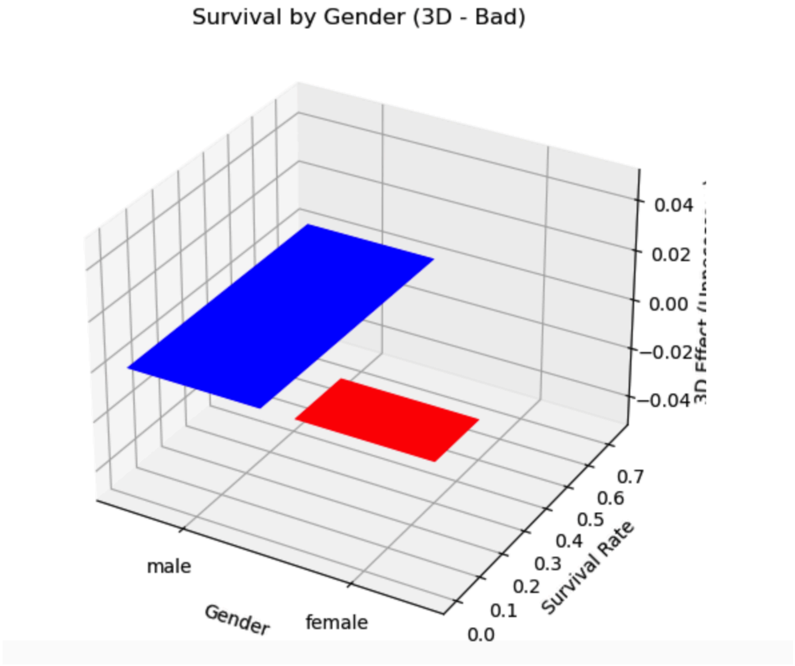


  - **Stacked Bar Chart:** Displayed gender distribution within passenger classes and survival outcomes, providing insights into both survival and class interaction. Shows that males were mostly third class, while females were more equally distributed.

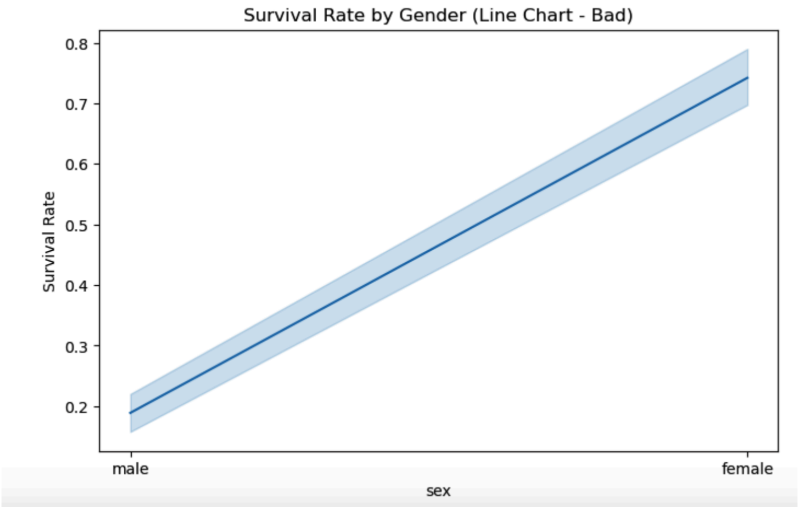


- **Ineffective Visualizations:**

- **3D Bar Chart:** Added unnecessary complexity and clutter, making it harder to interpret survival data by gender.



- **Line Chart:** Not suitable for categorical data, best used for continuous data and trends.



Overall, gender had a noticeable impact on survival rates, and the visualizations either effectively or ineffectively highlighted these differences. While some charts provided clear insights, others, like the 3D chart, added unnecessary complexity without enhancing understanding. Choosing the right visualization is key to presenting data clearly and accurately, especially when focusing on categorical comparisons like gender and survival.

Conclusion

Based on the comparisons above, one can see the importance of choosing an appropriate visualization technique when conveying information through graphics. The bar graphs provide easy to follow comparisons with appropriate labeling. On the other hand, the visualizations illustrating survival rates by gender and deck location are visually confusing making it hard, if not impossible to understand what is being displayed.

The right visualization provides the reader with an image that can be easily read and interpreted without additional context. This aligns with a key point made in the research article "Principles of Effective Data Visualization"[1] by Stephen R. Midway, specifically in section 8, titled "**Principle #8: Simple Visuals, Detailed Captions."** In this section, Midway writes, "Captions should be standalone, which means that if the figure and caption were looked at independent from the rest of the study, the major point(s) could still be understood."

The bar graphs provide easy to follow comparisons while the visualizations illustrating survival rates by gender and deck location are visually confusing making it hard, if not impossible to understand what is being displayed.

1. (Reference: Midway, S. R. (2020). *Principles of effective data visualization*. *Patterns, 1*(8), 100141. https://doi.org/10.1016/j.patter.2020.100141)