Visualization Ethics and Communication

Submitted By: Group 2

Course: DATA-6550

Introduction

Data visualization is among those powerful tools in a data scientist toolkit in which when used responsibly, it has complex known patterns to surface clearly and intuitively. In this project we used the Titanic dataset to draw analysis and conclusion on how visualization helps in decision making.

The sinking of the Titanic was indeed a disaster that happened in 1912 and it was one of history's most documented and well known tragedies. The current under study dataset tells about the passenger demographics and also it has information on survival outcomes, and it makes it an ideal candidate for exploring and finding different visualization design choices to either stand out or hide the truth. In this project we as a group of 3 members came up with 4 visualizations each among which two were designed to tell about the data and information it contains as accurately and clearly as possible, and the other two were intentionally designed to give false information or mislead the reader to demonstrate how the ethical issues can arise and how ethical rules come with datasets available online.

In this project we created this document which visualizes how each member produced different aspects of the data and how the reasoning and techniques behind each visualization works, and also tells how the different ethical rules are affected only by these visualizations.

Dataset Overview

The Titanic dataset is downloaded from Kaggle and it is open source data which contains the overall information on 891 passengers which are in the training split. We used the following features to carry out the analysis:

- Survived: Binary variable where 1 is for survived and 0 is for not survived
- pclass: Passenger class is a proxy for socioeconomic status and the dataset has 1st, 2nd and 3rd classes in it
- Sex: Passenger gender
- Age: Passenger age it has 177 missing values
- Fare: Ticket fare paid by each customer

In the training set the overall survival rate is around 38.4% in which 342 passengers were survivors among 891 total passengers. Then in the Age column missing values were filled using the median age, and in the dataset the two missing values were then replaced with the common value. All of these preprocessing steps also ensured us that the clean and reliable data for visualization.

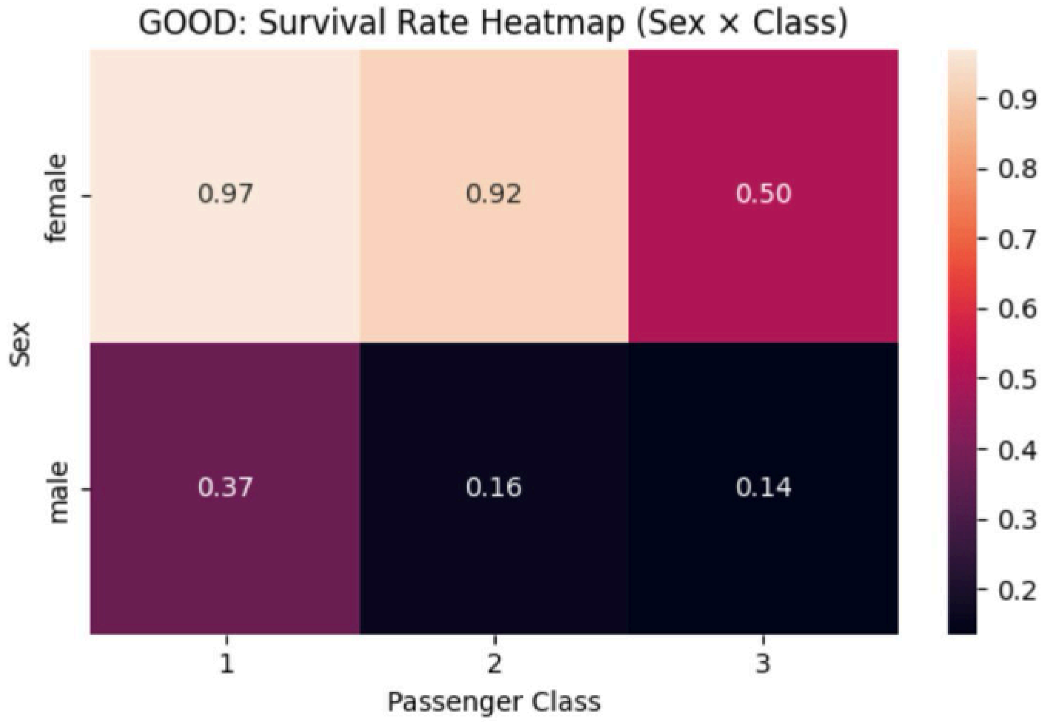
Individual Visualisation Contributions

Contribution by Momina L. Ali

Here are the following four plots which were created from the Titanic training dataset. Python libraries like seaborn and matplotlib were employed. Also, data preprocessing here included median values for missing Age and mode as wells.

Good Visualization 1: Survival Rate Heatmap (Sex × Passenger Class)

..



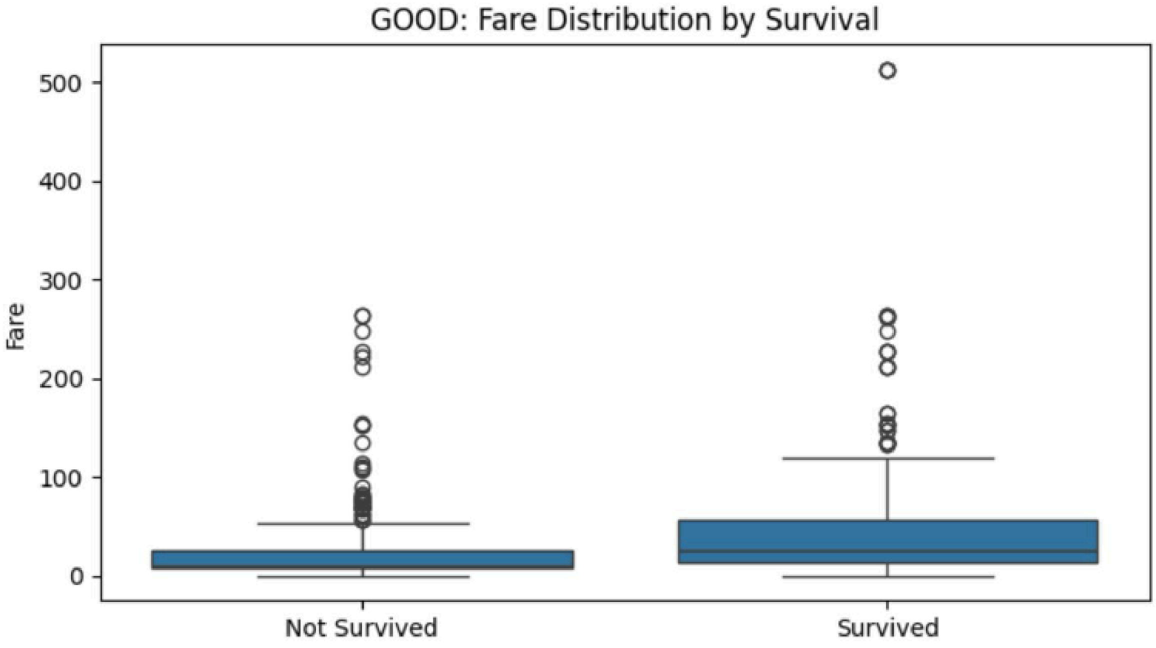
The above shown heatmap here presents the across six combinations of Sex and Passenger Class for survival rate. In the dataset, each of the cells contain the survival rate as an annotated value, and the color gradient from light to dark which represents low survival rate. This allows the readers to grasp the patterns immediately without even reading the numbers so they can see the patterns quickly.

Then this visualization is made strong because by showing both variables of class and sex show simultaneously and it shows the relationship between them and also tells that would be hidden if either variable were examined alone. For example, in the first-class cabins the females were 97% surviving, while the males who were in the third-class cabins survived at only 14%. The heatmap shows us how complex the data is, also it avoids random-picking, and uses an accurate color scheme to aid interpretation.

Good Visualization 2: Fare Distribution by Survival (Box Plot)

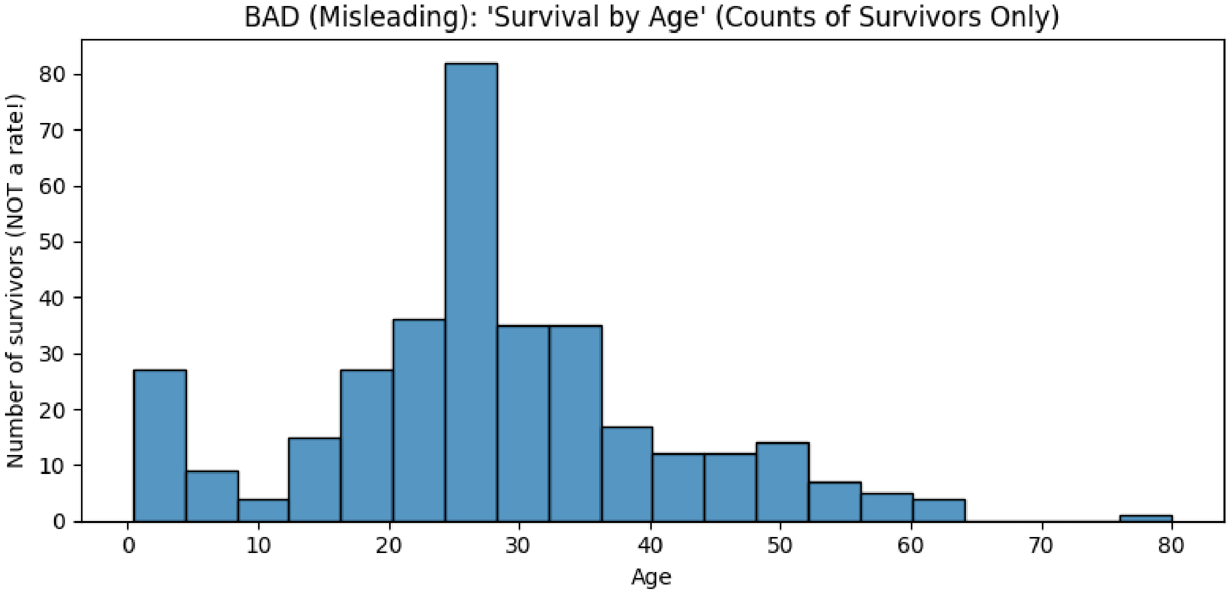
Now for another good visualisation we have considered a box plot which compares the distribution of overall price paid for each ticket by passengers who survived and those who did not survive. The following visualization shows us the median, interquartile range, and individual outlier points tell the viewers that the overall view of the price paid by passengers is distributed among each group and not among the single summary statistic.

The plot here tells us that those who survived in the disaster actually paid a higher fare: the average fare paid by passengers is higher in those who survived, and the IQR then extends to much higher in the upward trend as compared to those who could not survive. Now this tells us that the finding is informative and it tells the actual difference in the wealth or socioeconomic trend of our society in the survival rate of the Titanic incident, we can clearly see that the wealthy people who bought expensive tickets had higher chances of survival as compared to those who paid less to buy a ticket.



Misleading Visualization 1: “Survival by Age” Counts of Survivors Only

This histogram here shows that the age is directly related to the survival rate on the Titanic. But, it contains something that is critically deceiving. On the y-axis we have age and that is the raw count of those who survived but it does not tell the survival rate. This tells that which age group had more people who survived and simply they had more people to start with, and it does not always mean that they have a higher chance of surviving.



The totally deceiving direction may seem subtle but it is really significant. A reader who is looking at the plot will likely state or draw the conclusion that the people who were in their mid-20s had the most chance of survival, since that bin contains the most count. But in reality, the age range of 20 to 30 had the highest number passengers on the ship that day. The actual survival rate which is across age groups is more symmetric or in other terms uniform. In this example the frequency is being confused with the probability which is a common error in data visualization. To correctly infer the true information from this plot, we will divide the overall survivor count that is in each bin with the total number of passengers only in that specific bin to get the actual survival rate. The misleading design choice is just the label of y-axis which is ‘Number of survivors’, and for that most casual readers will not notice or interpret carefully.

Misleading Visualization 2: Average Fare by Sex (Ignoring Passenger Class)

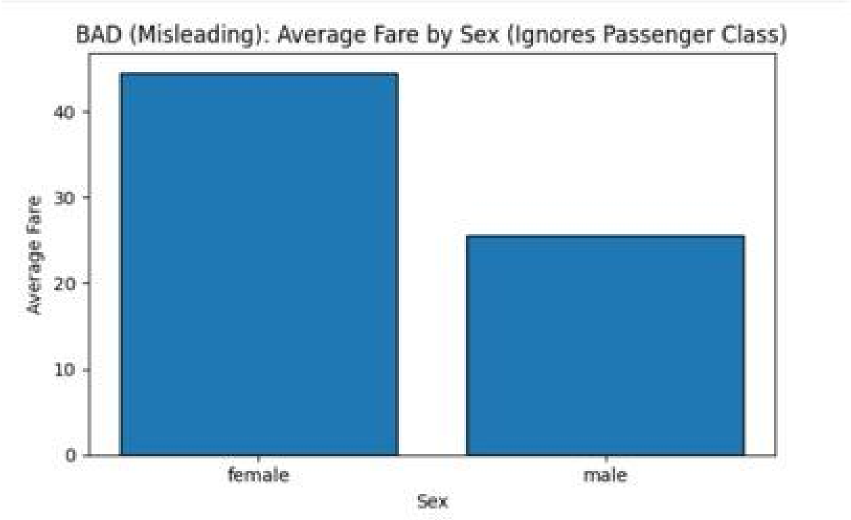


Figure 4 (Misleading). Bar chart showing average fare by gender, omitting the confounding effect of passenger class.

The above bar chart tells us that the females paid a higher ticket price than males which is around £44 vs. £26. If we look at the plot for the first time, this may seem to be saying that gender played a critical role in ticket price. But, this statement is totally misleading since the chart does not take into account the critical variable of Passenger Class.

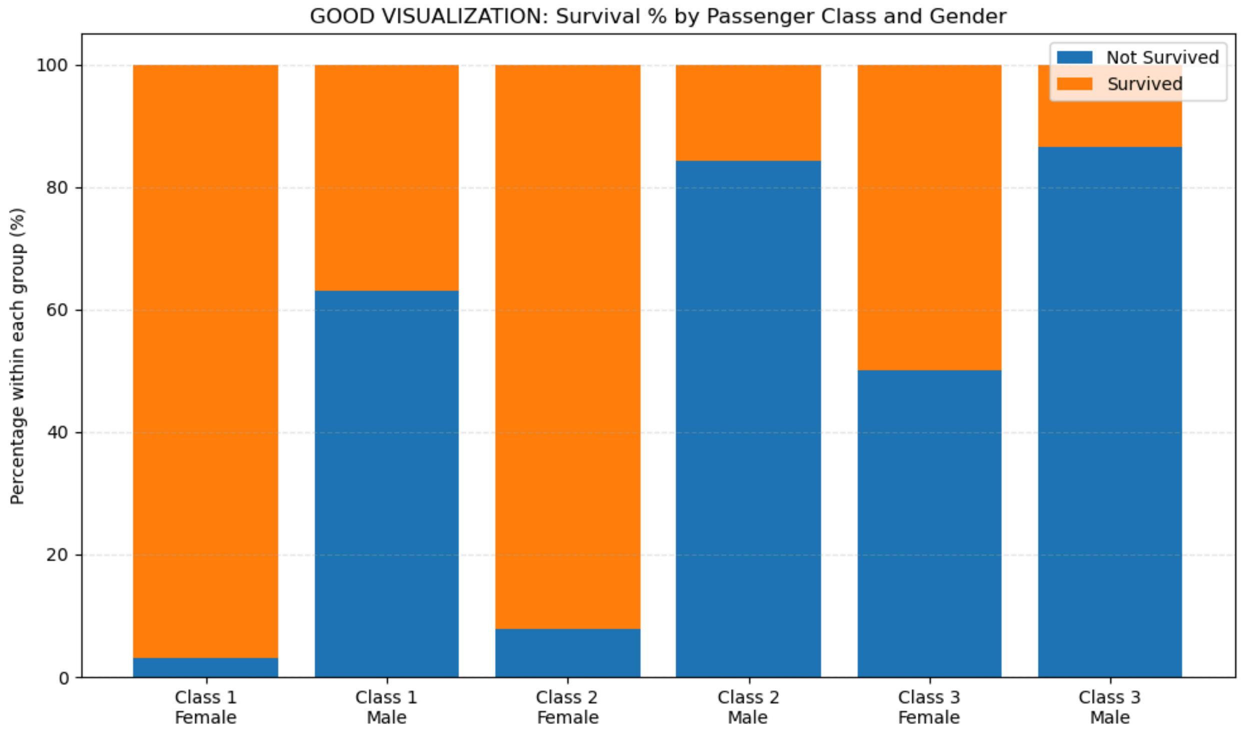
In reality the women passengers, especially in the first and second class, were dominating numbers as compared to third class in comparison with the male passenger distribution. First-class tickets were more expensive than third-class tickets. And because more females were in higher class, the average fare among all women seems to be inflated compared to men and this is not because of gender-based pricing or tickets were expensive for females, but it was because of an imbalance in class. The visualization shows exactly what the passenger class breakdown is and it presents an incomplete story that tells us a false causal relationship. A true version of this chart would break down the average price of ticket for both sex and class simultaneously which will make the overall confounding structure a lot more visible.

Contribution by Satvick Yadlapalli

Good Visualization 1: Survival % by Passenger class and Gender

This visualization works well because it presents the survival pattern in a very clear and structured way. The chart compares survival percentages across both passenger class and gender at the same time, which helps us understand how these two factors together influenced the outcome. Since each bar represents 100 percent of a specific group, we are looking at proportions instead of raw counts. That is important because the number of passengers in each class and gender group is different, and using percentages avoids misleading conclusions.

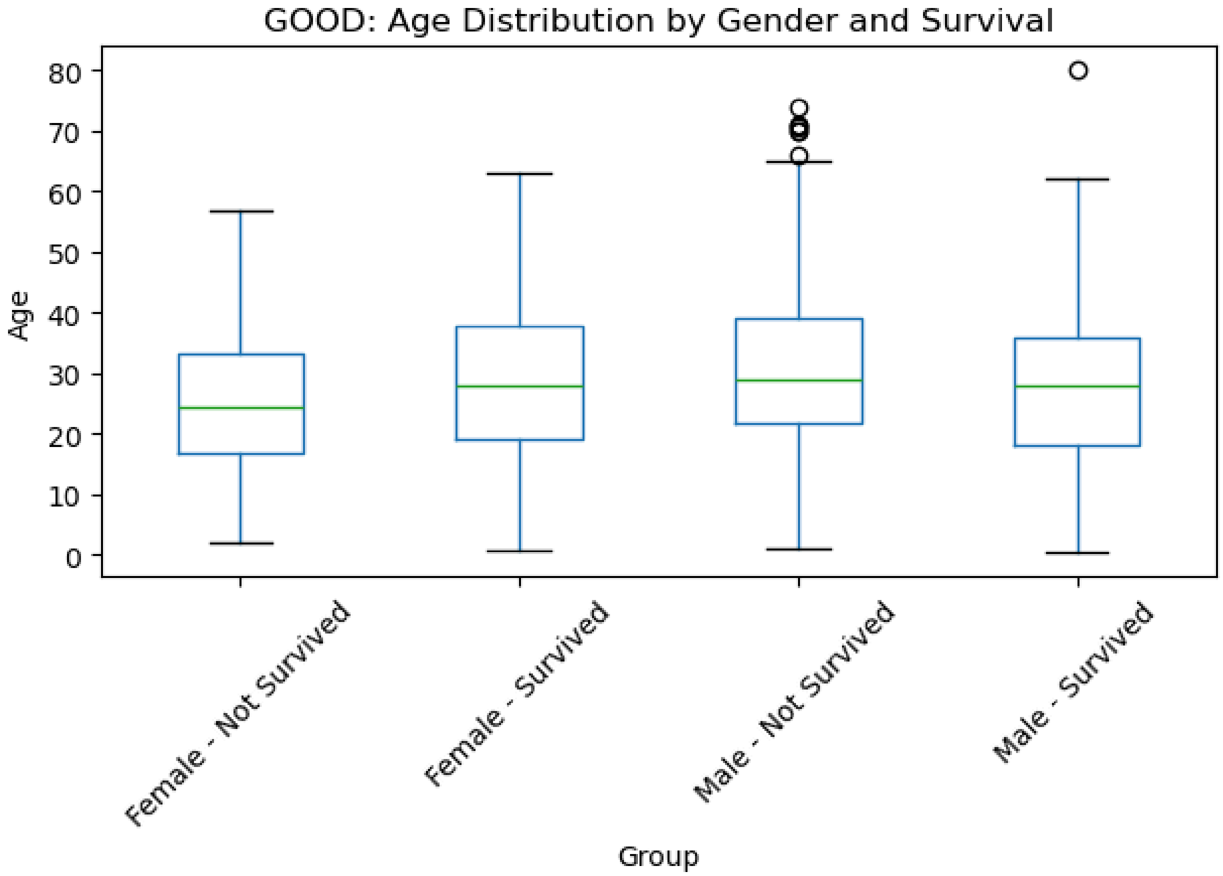
The stacked format also makes it easy to compare survival and non-survival within the same group. We can quickly see the difference between males and females in each class without doing any calculations. For example, it becomes obvious that females had a much higher survival rate than males in all classes. At the same time, we can observe that first-class passengers had better survival outcomes compared to third-class passengers. These patterns are visible immediately, which means the chart communicates insight effectively.The color choice is simple and consistent, using only two colors to represent survived and not survived. This keeps the design clean and prevents distraction. The labels on the x-axis clearly separate class and gender, the y-axis shows percentages, and the title directly explains what the chart is about. Because of this clarity, the viewer does not need extra explanation to understand the message.



Good Visualization 2: Age Distribution by Gender and Survival

This visualization is strong because it gives a complete picture of how age relates to survival without making the chart complicated. Instead of only showing the average age, the box plot shows the median, which is more reliable when the data has extreme values. By separating the groups into female and male, and then into survived and not survived, the chart allows us to compare patterns clearly within each category. We can quickly see whether the typical age of survivors is different from non-survivors and whether that pattern changes between genders. This makes the relationship between age and survival easier to understand in a structured way.

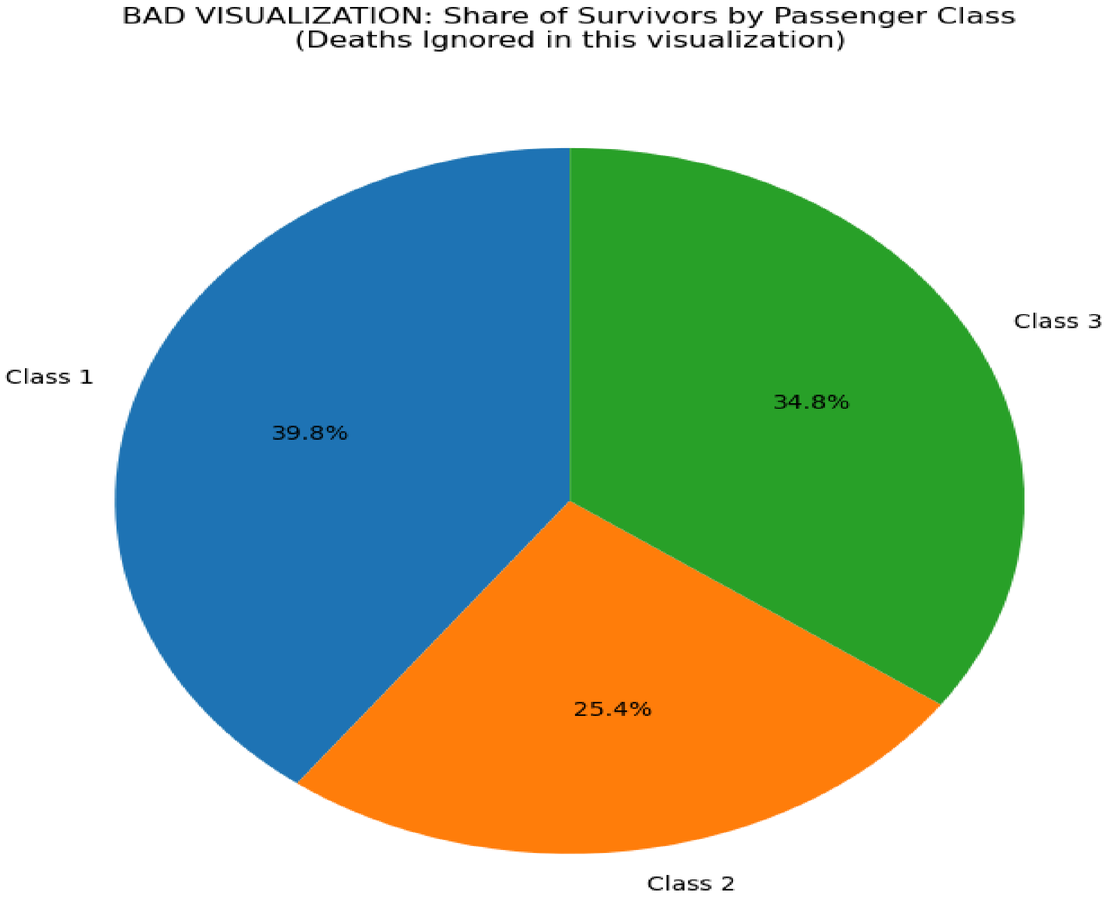
Another important strength of this visualization is that it shows variation, not just one number. The size of each box tells us how spread out the middle 50 percent of ages are, and the whiskers show the overall range. This helps us understand whether ages were concentrated in a small range or widely distributed. The outliers are also marked clearly, which is useful because extreme ages can affect interpretation. Since age is continuous numerical data, a box plot is a very appropriate choice because it summarizes distribution, central tendency, spread, and unusual values all in one figure. Overall, this chart provides deeper insight into the data while staying clean, readable, and easy to compare across groups.



Misleading Visualization 1: Share of saviors by Passenger class

This visualization is misleading because it only shows the share of survivors from each passenger class, without considering the total number of passengers in each class. By displaying only survivors in a pie chart, it hides the actual survival rate. A viewer might assume that Class 3 performed well simply because it takes up a large portion of the pie. However, this chart does not show how many total passengers were in Class 3 or how many of them did not survive. Since Class 3 had the largest number of passengers overall, it naturally also had a noticeable number of survivors, but that does not mean their survival rate was high.

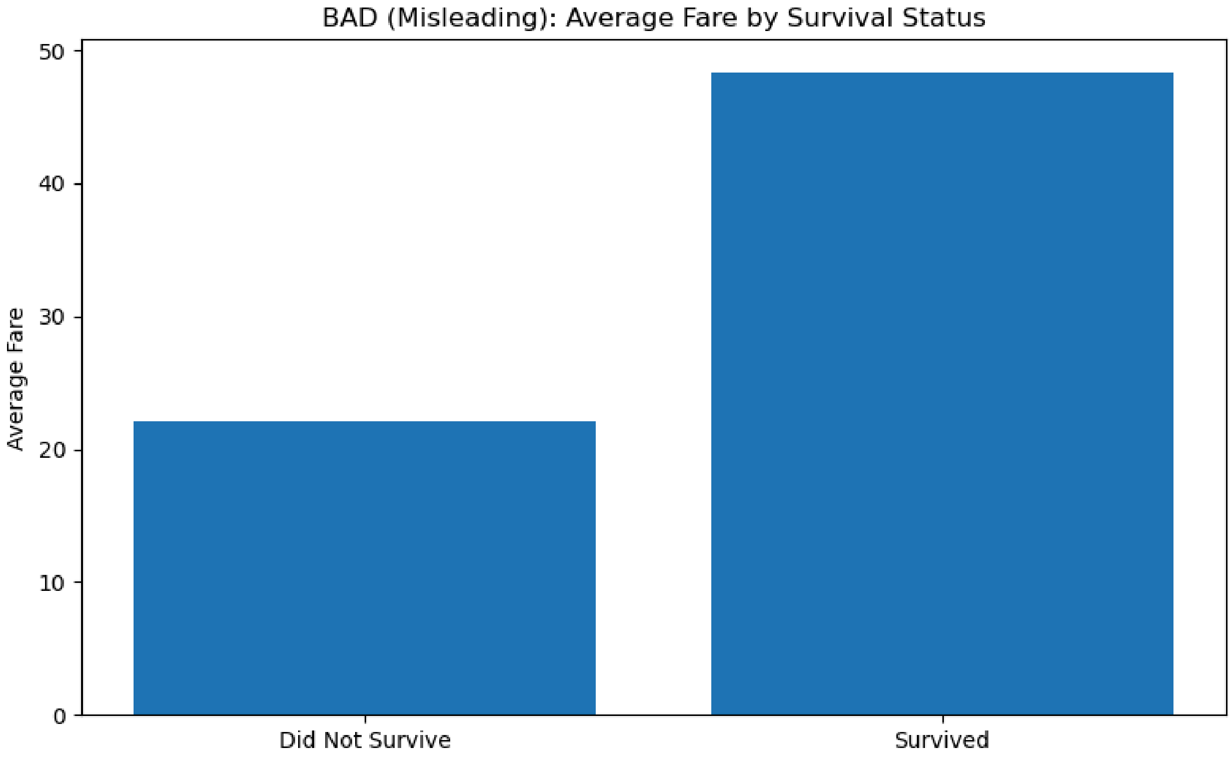
The main problem is that the chart focuses only on proportions of survivors instead of survival percentages within each class. This can easily confuse someone who is not carefully analyzing the data. In reality, the survival rate for Class 3 was much lower compared to Class 1, but this information is completely hidden here. By ignoring deaths and total counts, the chart gives an incomplete picture and may lead to incorrect conclusions. A better visualization would compare survival rates within each class rather than just showing the share of survivors.



Misleading Visualization 2: Share of saviors by Passenger class

This visualization is misleading because it uses the average fare instead of the median. The Titanic dataset has some passengers who paid extremely high fares, especially in first class. These few very high values can pull the average up and make it look much larger than what most people actually paid. So when we see that survivors have a much higher average fare, we might assume that paying more almost guarantees survival. In reality, that conclusion may not be accurate because the mean is being influenced by a small number of very wealthy passengers.

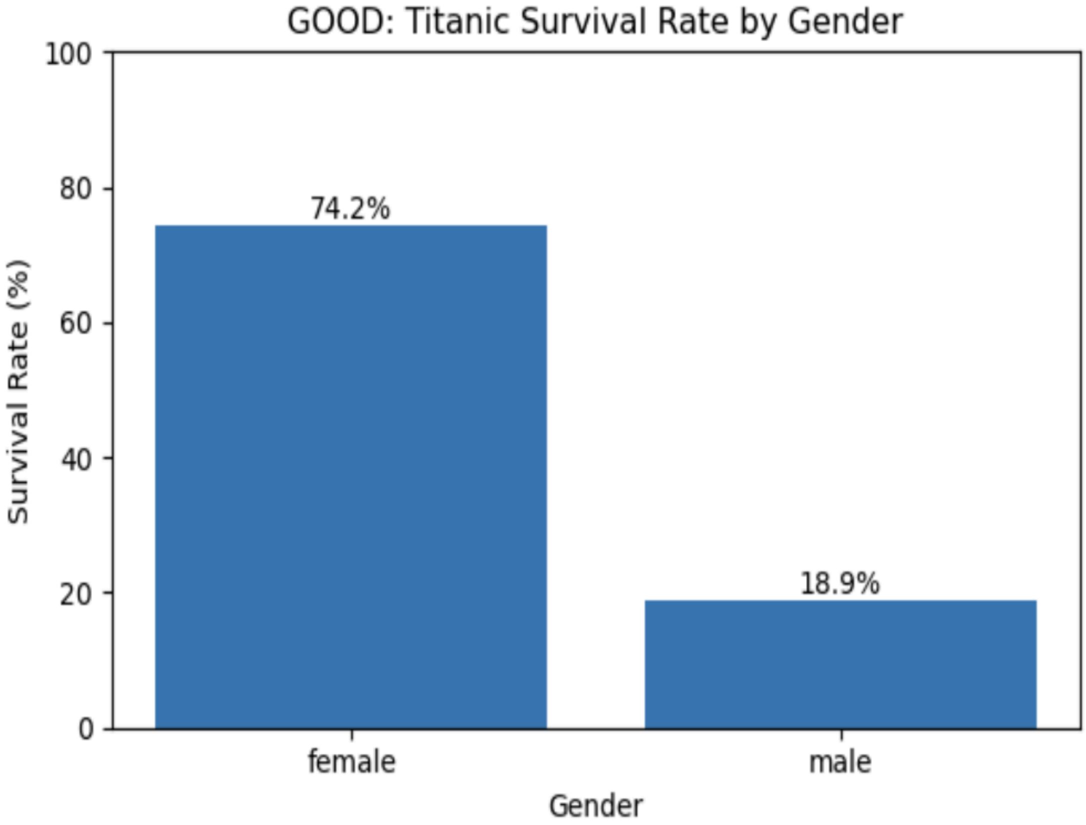
Another issue is that this chart does not show how the fares are distributed. We cannot see whether most survivors paid similar fares or if the data is spread out widely. It only gives one summary number for each group, which hides important details like variation and outliers. Because of this, the chart oversimplifies the relationship between fare and survival and may make it seem stronger than it actually is. A better visualization would show the distribution, such as with a box plot, so we can understand the full picture instead of just relying on averages.



Contribution by Suchitra Hallikeri

As I contributed to this project, I focused on the creation of visualizations that effectively demonstrate the impact of minor design decisions on interpretation. Through the utilization of Python libraries such as pandas and matplotlib, I was able to create two ethically sound visualizations that accurately depict survival trends within the Titanic dataset and two misleading visualizations that can be easily manipulated for such purposes

Good Visualization 1: Survival Rate by Gender



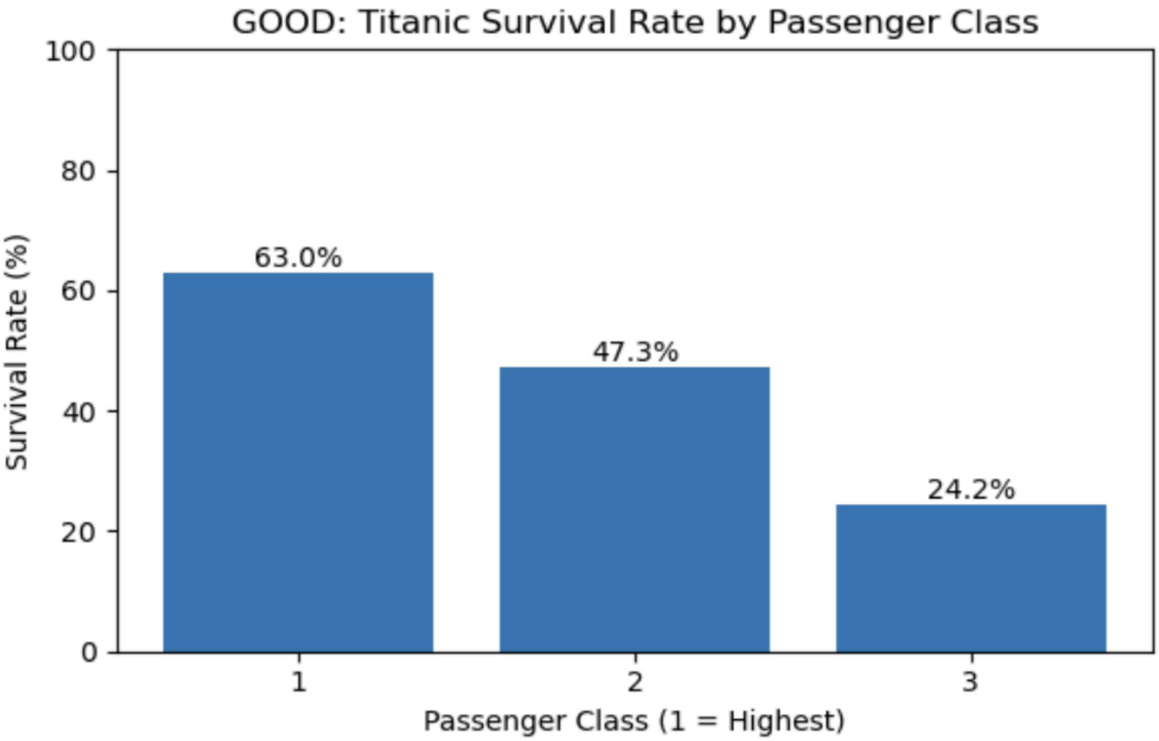
The bar chart provided depicts the survival rate of both male and female passengers on the Titanic using the full training dataset. The chart does not display the actual number of male and female passengers but displays the survival rate in terms of the percentage of the total population. The display of the survival rate in terms of the percentage of the total population helps in making a fair and proportional comparison between the survival rate of both male and female passengers.

The bar chart indicates that there was a notable difference in the survival rate of male and female passengers. The survival rate of female passengers was about 74.2%, while that of male passengers was only about 18.9%. The difference in the survival rate can be explained by the fact that women were given priority in accessing the lifeboats.

From an ethical perspective, the bar chart provided can be considered effective in several ways. First, the y-axis starts at zero, thus making the height of each bar an accurate representation of the actual difference in the survival rate of both male and female passengers. The display of the survival rate in terms of the percentage of the total population helps in avoiding any exaggeration that could occur if the actual numbers were used in the bar chart. The display of the survival rate in terms of the percentage of the total population helps in avoiding any exaggeration that could occur if the actual numbers were used in the bar chart.

The display of the survival rate in terms of the percentage of the total population helps in avoiding any exaggeration that could occur if the actual numbers were used in the bar chart. The display of the survival rate in terms of the percentage of the total population helps in avoiding any exaggeration that could occur if the actual numbers were used in the bar chart.

Good Visualization 2: Survival Rate by Passenger Class

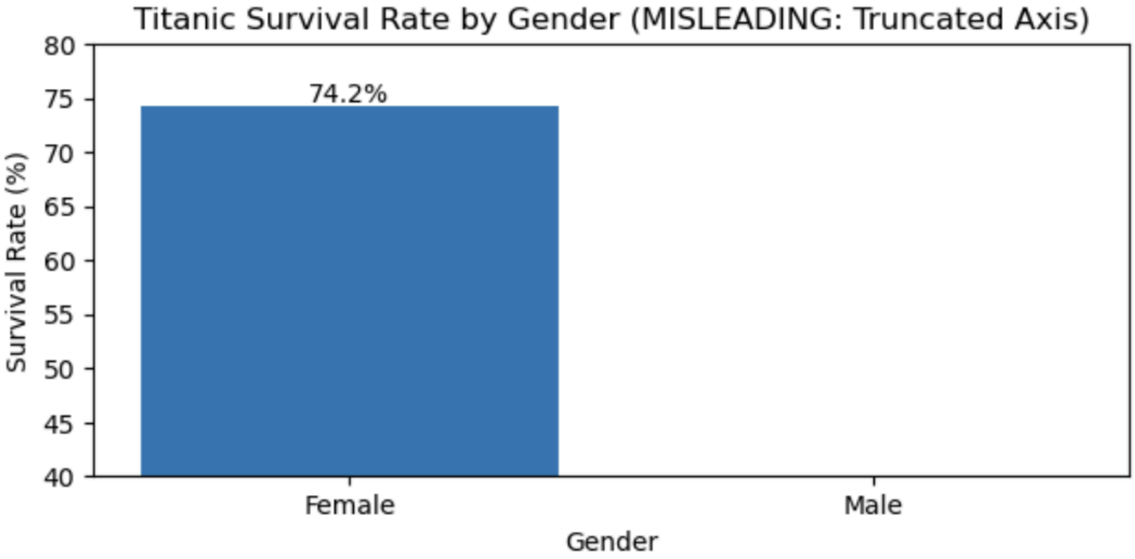


This bar chart shows the survival rates for male and female passengers on the Titanic using the entire training dataset. Instead of using absolute numbers, the chart shows survival rates as a percentage, making it possible to compare the survival rates for both men and women in a fair and proportional manner despite the absolute numbers for each group differing significantly.

From the chart, it is possible to note a stark difference in survival rates for men and women, with a survival rate of 74.2% for women compared to 18.9% for men. This is consistent with historical evidence on the Titanic disaster, in which women were given priority in boarding lifeboats compared to men.

Ethical Considerations
This visualization shows excellent ethical consideration in several ways:
First, the chart uses a zero-based scale for the y-axis, which makes it possible for the relative heights of the bars to accurately represent the data without room for distortion through exaggeration, especially because bar charts are based on relative bar height for data comparison.
Second, using a percentage-based chart reduces the likelihood of false conclusions arising from differences in absolute numbers.
Third, the chart is minimalistic in its design, making it possible for the data to speak for itself with minimal external influence through distractions. This chart shows excellent ethical consideration in its ability to accurately represent the data for a fair and proportional comparison of survival rates for both men and women despite the absolute numbers for each group differing significantly.

Misleading Visualization 1: Survival Rate by Gender with Truncated Y - Axis

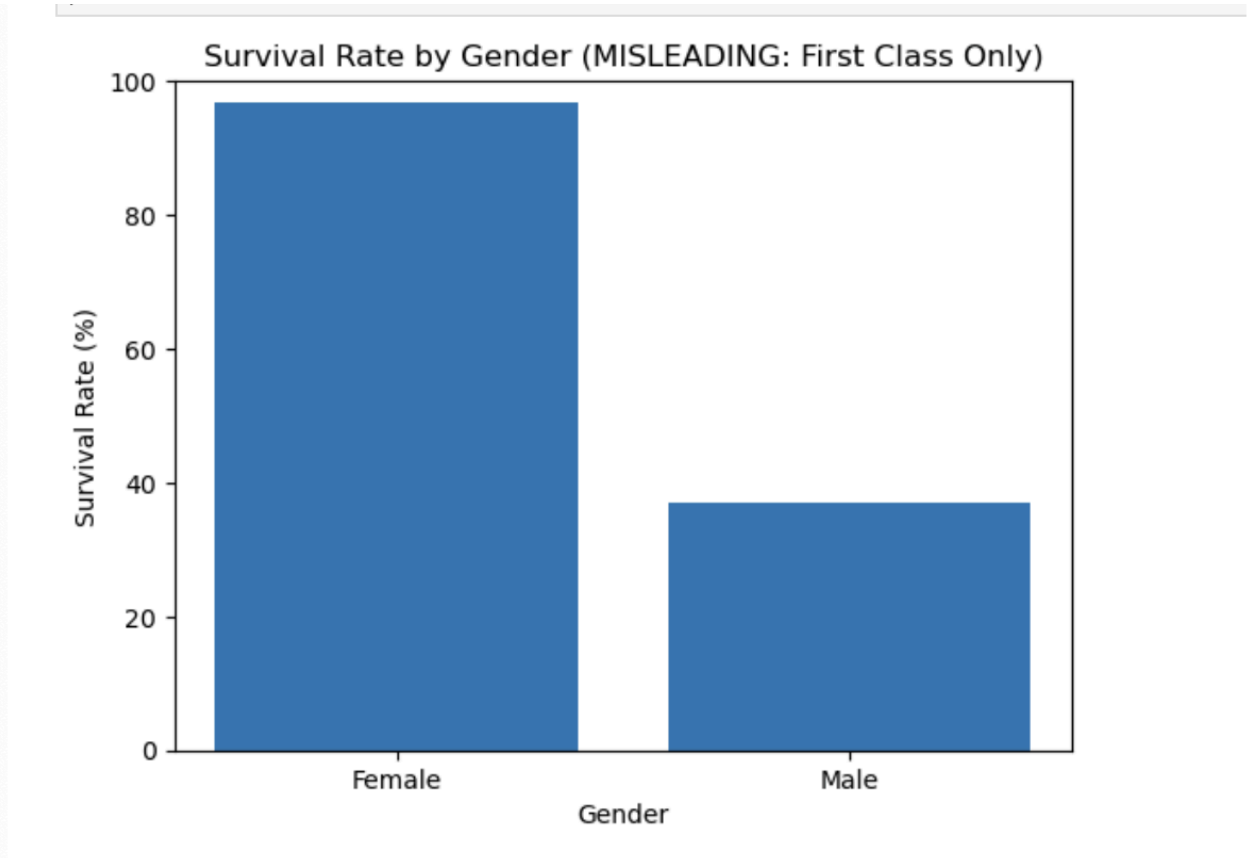


This bar chart shows information regarding survival rates for male and female passengers, but it intentionally truncates the y-axis, which starts from 40% instead of zero. The information provided in the chart is correct, as female passengers have a survival rate of 74.2%, and male passengers have a survival rate of 18.9%. This chart, however, intentionally exaggerates the difference between male and female survival rates by truncating the y-axis.

A bar chart mainly uses height to represent data. When a bar chart truncates the y-axis and starts from a number other than zero, even a small difference between two values can be significantly larger. This bar chart shows that female passengers have a much higher survival rate, and their bar is much taller, whereas male passengers have a survival rate that is barely visible in this chart.

This bar chart is misleading because it violates a fundamental principle of a bar chart, which is that it must start from zero. This is not a problem with the data, but it is a problem with how it is presented. This is a misleading bar chart because it exaggerates the difference between male and female survival rates, which is not a correct representation. The correct bar chart would be one that extends from zero to 100% instead of 40% to 100%.

Misleading Visualization 2: Survival Rate by Gender (First-Class Passengers Only)



The bar chart focuses on survival rates based on gender; however, it only presents information for first-class passengers. Based on a quick glance at this bar chart, one might get an impression that survival rates for those on board the Titanic were quite high. In addition, survival rates for women might be close to 100%. However, it should be noted that survival rates for second- and third-class passengers were lower than those for first-class passengers.

The information provided in the bar chart is accurate; however, it might be argued that it does not provide a complete picture. The information provided in the bar chart focuses on survival rates for one class of passengers. As such, it might be argued that it presents a biased view of survival rates for those on board the Titanic. In addition, it does not present a complete picture to the audience. Survival rates for all classes of passengers would be needed to present a complete picture. In addition, it would be important to compare survival rates for all classes of passengers.

Why This Is Misleading

The visualizations in this section show the ease with which perception can be altered without changing the underlying data. In the case of the truncated axis, the actual values are correct. However, the limited y-axis gives the false impression of greater difference between the groups. It should be understood that humans naturally compare values based on the height of the bar. Therefore, the altered axis gives a false impression of the actual values.

The first-class-only visualization is also false in the sense that it omits crucial information. It omits the information pertaining to the passengers in the second and third classes, who faced much lower survival rates. Although the actual figures in the visualization are correct, the omission of crucial information gives a false impression.

The above two visualizations show that the bias in the visualization of the data might not be due to the actual figures but due to the context in which the figures are presented.

Final Ethical Reflection

The objective of this project is to continue to validate the idea that data visualization can be an effective form of storytelling. The presentation and contextualization of information has the ability to shape and change beliefs, opinions, and even decision-making strategies. Even trivial factors such as how one scales a graph or what information to include or exclude can greatly impact how one interprets a certain set of information.

Through an examination of both honest and dishonest data visualizations, it was made clear how easily a form of visual distortion can be created, and how important it is to be transparent, complete, and proportionate in data visualization. Honest data visualization is a process that is carefully considered, where simplicity is not necessarily a priority, but where context is maintained.

As future professionals in the world of data, it is understood that a data visualization is not necessarily objective. Each data visualization tells a story, and it is up to us to ensure that it is a story that accurately represents the entire data set, not a carefully constructed subset thereof.

Ethical Considerations in Data Visualization

The visualizations that are made for this project show several ethical drawbacks that data practitioners must take care of. From our group, four key themes were drawn.

Counts vs. Rates

A common deceptive practice is presenting raw counts as proportions or rates. For example, Misleading Visualization 1 plots the number of survivors by age, which suggests certain age groups were more likely to survive. In reality, the chart reflects the demographic makeup of the passenger manifest. Visualizations that compare groups of different sizes without normalizing the data can create similar distortions.

Omitted Variables and Confounding

Misleading Visualization which is plot 2 shows how extracting a confounding variable can make something by itself kind of information and apparent relationships. Telling average fare is based on gender and no control on passenger class then creates the totally false perspective that gender was one of ticket price factors. But in reality, the passenger class which has a direct link with gender in this dataset is the actual factor. Ethical visualization wants us as the analysts to take into account where it is possible to display all features that meaningfully affect the outcome being studied.

Choosing the Right Chart Type

Good Visualization plot 2 shows us that the chart type selection is also an ethical and well aesthetic decision. When we use a box plot and not the simple bar chart of averages, then we see a skewed fare distribution and then we have outliers as well which are fully visible to the reader. Dividing this whole information into a single average bar will be a technically accurate decision but it will make the picture incomplete. Then by selecting chart types that hide variance also that truncate axes can mislead the reader even without throwing any false data into the plot.

Multi-Dimensional Representation

Good Visualization 1 also tells us the value of showing data in different dimensions at a time. When we plot both the sex and passenger class together in a heatmap then, the visualization helps the reader to only extract the oversimplified single-variable conclusions. Ethical data visualization also sometimes requires going away from the simplest chart and putting our effort in the extra design so we can represent the complete complexity of the data.

Conclusion

This project of visualisation gives us a core lesson of data science which is that the way data is visualized actually directs the audience's perspective and instills thinking among them and shapes their belief. The Titanic dataset, with its variables of class, gender, age, and fare, offered a good chance to both illuminate and obscure the truth depending on the choices made.

Our group came up with accurate visualizations which give priority to completeness and also appropriate chart selection, and also represent the distribution and interaction effects. Our misleading visualizations or bad visualisations shown in this report exploited common ways our cognitive systems perform in which confusing counts for survival rates, and drawing conclusions from incomplete data to show how easily audiences can be deceiving even by plots that hold no outright wrong values.