Titanic Visualization Ethics

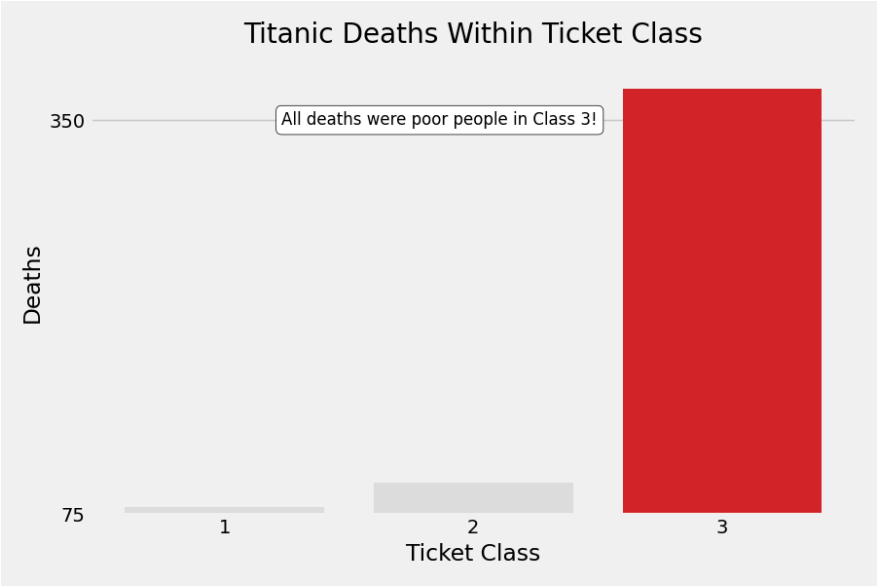
Group 1
Christian Adcock | Daniel Guthrie | Donovan Manogue | Ethan Stanks

Every day, millions of people come face-to-face with data, but most are unaware of it. Data is cleaned, processed, and engineered to create meaning. That meaning is then expressed by transforming relationships and trends in the data, into visualizations. Creating accurate data visualizations can help make various types of audiences aware, informed, and gain insights into ongoing problems and topics. However, not all visualizations tell the truth. Many authors of data visualizations attempt to mislead their audience for a specific goal in mind. It's the audience that needs to be informed about identifying those who try to mislead them through data.

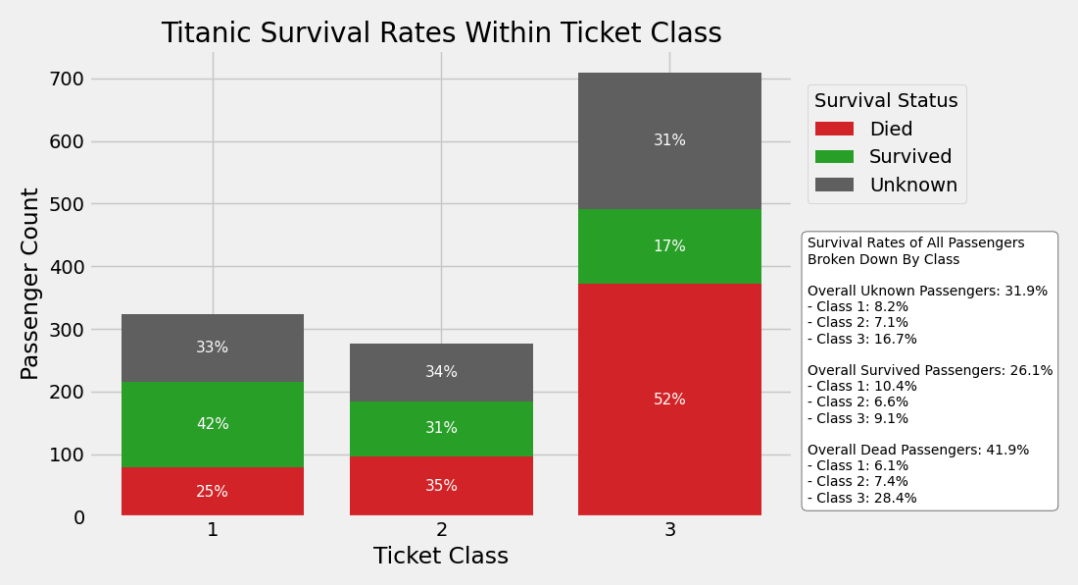
In this report, Titanic data, from one of history's most famous shipwrecks, is used to produce accurate and misleading visualizations. Many people assume surviving a shipwreck is luck, but data can be used to explore how possible it is for a passenger to survive. The dataset itself was collected from Kaggle, Titanic - Machine Learning from Disaster. It came compressed, split into a train and test dataset. Visualizations created from this data are from combining the train/test split or from the individual sets themselves. Each record in the data represents a passenger on the Titanic. The data contains ten columns: survival, pclass, sex, age, sibsp, parch, ticket, fare, cabin, and embarked. Survival represents if the passenger was recorded as surviving the shipwreck. Pclass represents what ticket class the passenger was in: 1 for 1st class, 2 for 2nd class, and 3 for 3rd class. First class tickets were more expensive and had a more luxurious lifestyle onboard. Third class was cheaper, had rooms closer to the bottom of the boat, and packed a lot more passengers into smaller spaces. Sex denotes whether the passenger was male or female. Age denotes how old the passenger was at the time. Sibsp marks the number of siblings and spouses the passenger had aboard. Parch marks the number of parents and children the passenger had aboard. Fare represents how much the passenger

paid for their ticket. Cabin labels the exact room number the passenger stayed in on the ship. Lastly, embarked represents the port from where that passenger boarded the ship; C for Cherbourg, France, Q for Queenstown, Ireland, and S for Southampton, United Kingdom. With all this data, many different data scientists can accurately tell a story of the odds a passenger had while aboard the Titanic. Nevertheless, some scientists can spin a tale and misinform their audience about that historical day.

A popular opinion about the Titanic is that passengers who were staying in a lower cabin on the ship had a less likely chance of surviving. In the movie “Titanic” featuring Leonardo DiCaprio as Jack Dawson, we see a man from third class live the life of a passenger in first class. The movie depicted those in third class to be poor, sharing the same living spaces, and locked behind a door as the ship goes down. How accurate was the movie compared to real life? As an average moviegoer, a data scientist, or an average Joe could watch that film and denote that all deaths on the Titanic were poor people in Class 3. That person loads up the Titanic dataset and begins making visualizations to prove to others that the movie is just like real life. A visualization is created by filtering the data to those that had their survival marked as False (dead) for each ticket class.



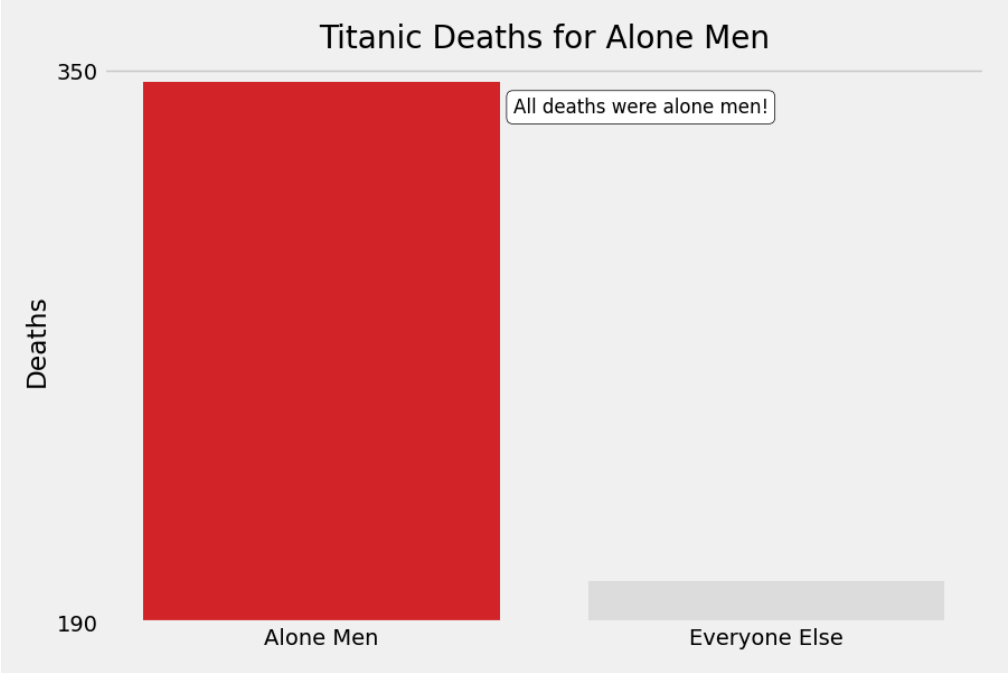
The visualization, “Titanic Deaths Within Ticket Class”, does exactly that. It conveys that almost all of those who died on the Titanic were passengers from ticket class three. Ticket class three’s bar towers over the other classes with a value above 350 dead. The other classes are very short compared to class three which means they had fewer deaths, right? Class three’s bar is colored in red to make those 350 dead dramatic compared to the others. This data scientist’s goal was to align the data to their opinion. However, that graph is very misleading because of that. The y-axis starts at 75 passengers, almost cutting off class one completely. Classes one and two’s bars are greyed out as if they don’t matter. A note puts a thought into the audience’s mind, forcing an opinion onto the visual. Lastly, it hides the fact that the majority of the survival rates are missing for each class. Drastically skewing the results. An accurate visualization would represent those who survived, died, and passengers who have an unknown result.



The visualization, “Titanic Survival Rates Within Ticket Class”, is an accurate representation of every passenger’s survival status for each ticket class. This visualization shows the full spread of passenger count from 0 to a little over 700. Each bar is stacked and color-coded to represent the survival status of passengers in that specific ticket class. There is no information hiding from the

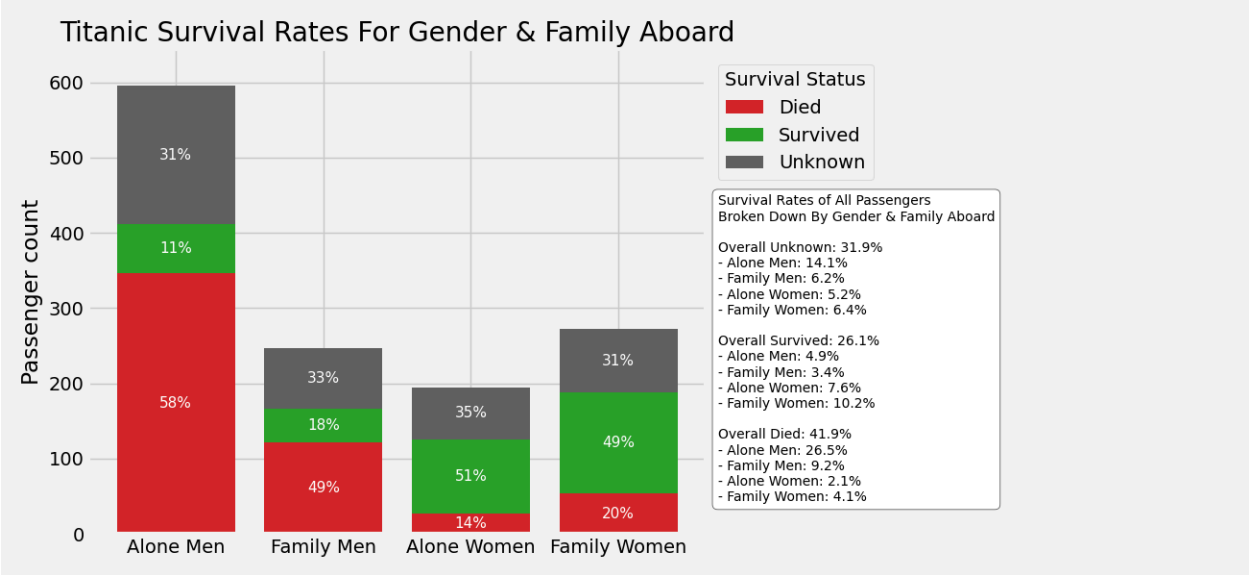
audience. Percentages are given to accurately convey how many passengers in that specific class have that survival status. For instance, ticket class 3's 'Died' stacked bar has 52%. This means 52% of all passengers in ticket class 3 did not survive the Titanic. On the right hand side of the graph there is an info box. The info box contains the overall breakdown of the survival rates, broken down by each ticket class. For instance, out of all passengers, only 26.1% of them survived. While 9.1% of those 26.1% were in ticket class 3. This visualization showcases that ticket class 3 had an abundant amount of passengers compared to the other classes.

Going back to the movie, Jack didn't survive, even though he was a male hanging around with the upper class. Rose, played by Kate Winslet, was a woman in first class who survived at the end of the movie. The movie showed that when the ship started to sink many of the crew prioritized women and children. A data scientist sees Jack, a male who boarded the ship alone, and assumes all men who were alone on the ship had to have the worst odds at surviving.



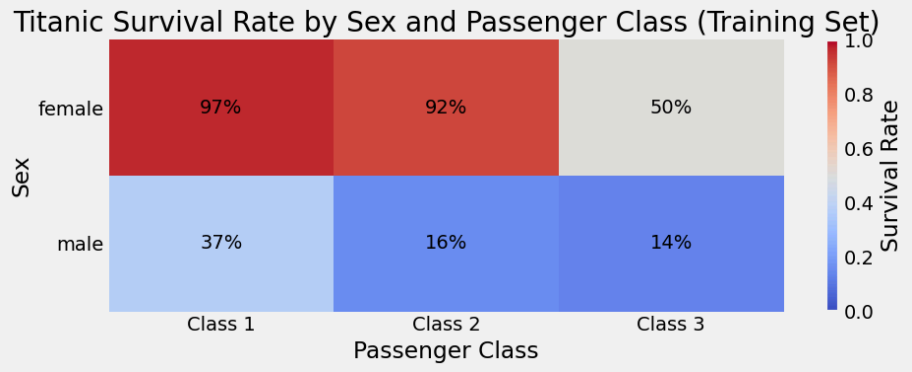
The visualization, "Titanic Deaths for Alone Men", did just that. It tells the audience that men had a towering defeat compared to everyone else aboard. Standing out in red, the bar representing alone

men, almost reaches 350 total deaths. While the small greyed-out bar, representing everyone else, comes nowhere near that 350 total. The visualization even includes a note, to spread an opinion, that all the deaths were men on board alone. How likely is it that even women alone all over the ship outlived a man? What makes a family man different from a man with no loved ones? This is where an accurate representation of the data is needed.



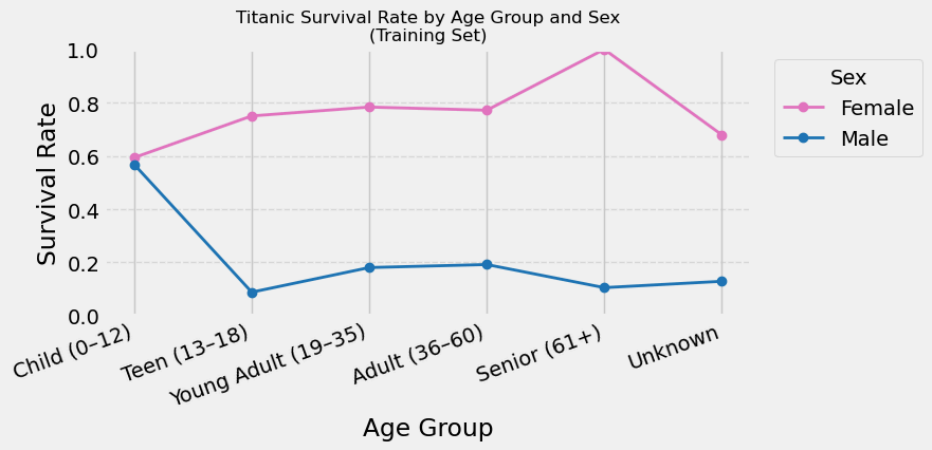
The visualization, “Titanic Survival Rates For Gender & Family Aboard”, accurately presents the populations of gender while grouping the two genders by those they’re onboard with. Unlike the previous graph, this one showcases men and women. Where family means those traveling with 1 or more other passengers marked as siblings, spouses, children, or parents. Each stacked bar represents the survival statuses of the passengers in the four groups. This graph does not hide the fact that 58% of single men died. However, it shows that 58% of alone men outnumber any other subgroup’s population entirely. This visualization highlights that nearly 50% of single and family women survived, while single and family men had less than 20% survival. An accurate visualization should show the entire picture, just like this one where it accounts for missing data. There is an info box on the right side of the graph that highlights the overall survival rates broken down into the four

groups. Taking a closer look, reveals that 14% of all unknown survival statuses are alone men. That could skew the survival results of those individuals. It accurately depicts what took place that day when the Titanic sank.

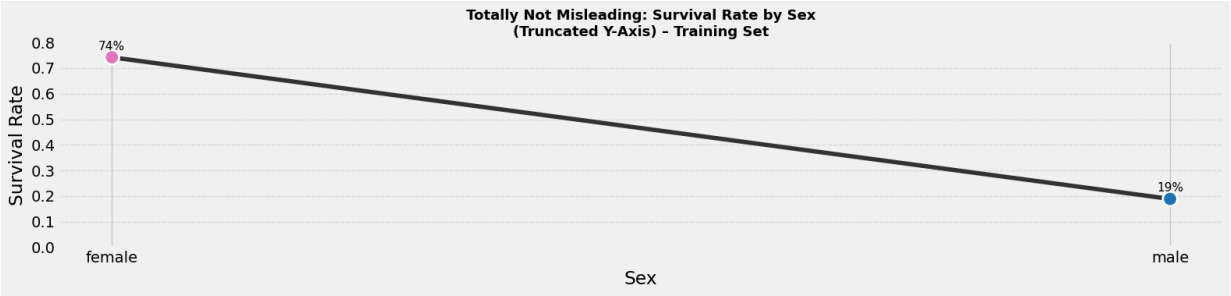


The visualization "Titanic Survival Rate by Sex and Passenger Class (Training Set)" is one of the two visualizations that accurately and effectively represent the data. It is important to note that this visualization is explicitly based on the training set data, meaning the patterns shown reflect only the subset of passengers used for model development rather than the entire Titanic dataset, and is made aware within the visualization's title to help with any confusion, being transparent and helping avoid any misleading or distorting any understanding. The heatmap of survival rate by sex and passenger class is accurate and effective because it presents the data in a normalized way rather than using raw counts. The graph accomplishes this by using the mean of the survival variable, which correctly communicates the probability of survival rather than just the frequency of survival. Doing this helps avoid misleading conclusions that could arise from differences in the size of each group. The color scale on the side ranges from 0 to 1, helping ensure that the graph is interpretable and consistent across all groups. Each cell is also labeled with its percentage, allowing viewers to pinpoint exact values without relying solely on color intensity. This layout makes it easy to compare both passenger class and sex survival rates simultaneously and reveals strong key patterns, such as the survival advantage for women in first and second class versus the sharp drop in survival for

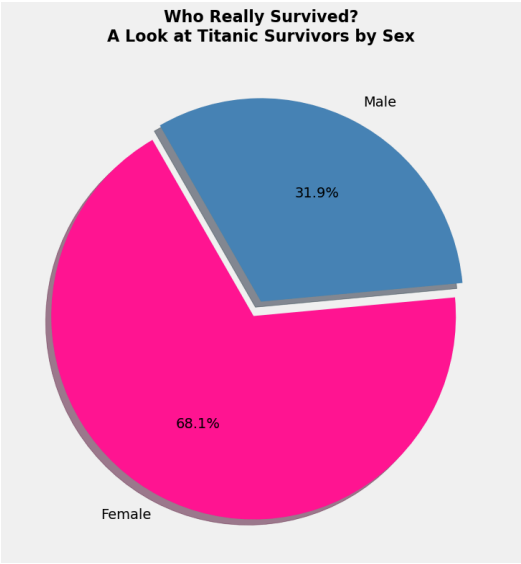
third-class passengers. Since the axis scales are not truncated or manipulated and the color scale maps the full range of possible survival rates, this visualization strives to be both transparent and analytically sound.



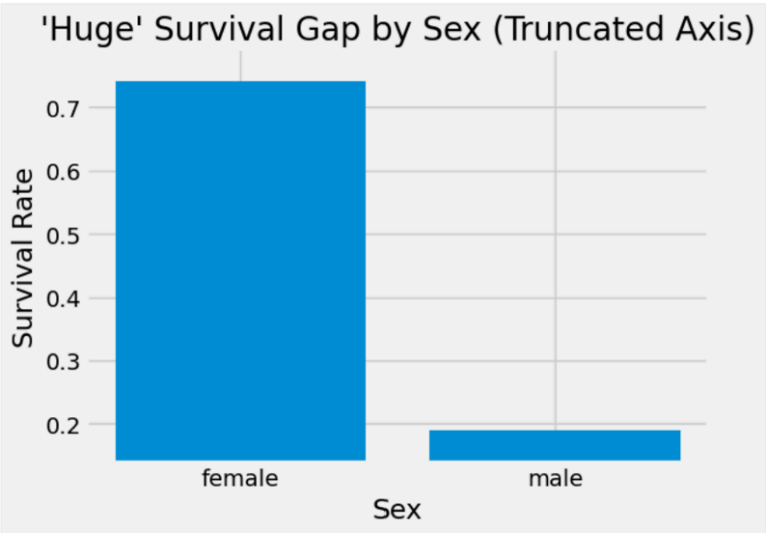
The second visualization that accurately and effectively represents the data is the "Titanic Survival Rate by Age Group and Sex (Training Set)".Again, it is important to note that the visual shows this is the training set, reflecting patterns from only that subset of passengers used in the model rather than the entire set. This second visual uses a line chart to capture the survival rate by age group and sex. It is accurate and very effective, as it preserves proportional comparisons across all age categories. The y-axis starts at 0 and goes to 1 (100%), maintaining honesty by showing the full range of differences in the graph. The consistent spacing of each group and the colors used to distinguish male and female passengers make any pattern easy to distinguish. This chart effectively communicates the survival differences not just by gender, but also by life stage. This graph shows that females consistently had a higher survival rate in nearly every age group. Since this visualization uses the entire training dataset, we included an unknown category to highlight any missed ages and avoid manipulation or exaggeration in the visual, and to clearly depict any trends or analysis. This helps the "Titanic Survival Rate by Age Group and Sex (Training Set)" visualization communicate the trends without distortion.



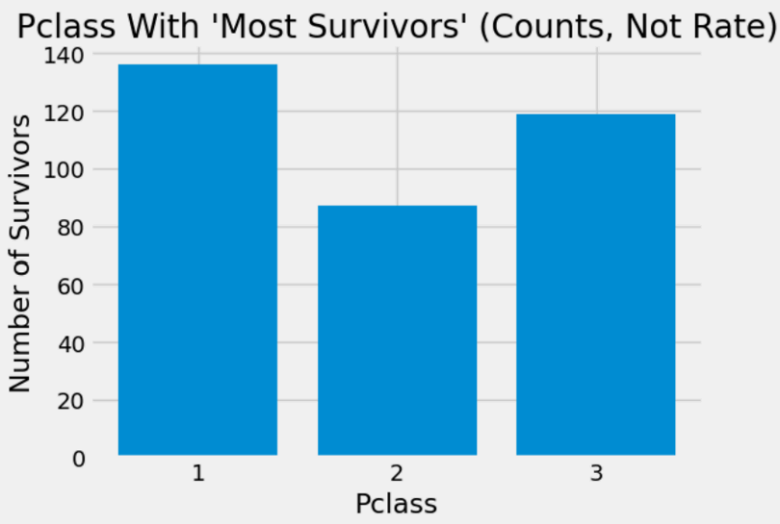
The third visualization, "Totally Not Misleading: Survival Rate by Sex (Truncated Y-Axis)", is the first graph on our list to be misleading. This visualization demonstrates how subtle design choices can distort perception. This visualization uses the correct survival rate calculations, but the truncated y-axis is cropping the range at 0.8 instead of using the full range of 0-1. It also exaggerates a massive visual difference with a steep slope. This can lead viewers to interpret the gap's magnitude as even bigger than it is. It is a big gap, but looking at this, it can definitely amplify the contrast visually. With the visuals' underlying numbers accurate, the axis manipulation and the y-axis cutoff can be manually controlled to determine how a viewer interprets the magnitude of the gap. This visualization highlights how specific scaling decisions can influence conclusions from the views without altering the data. Making accurate data look more dramatic than it already is.



The visualization, "Who Really Survived? A Look at Titanic Survivors by Sex," is something you would 100% see on Facebook, reposted by someone. This pie chart is misleading because it uses raw survivor counts rather than survival rates. While the percentages are correct, they represent the proportion of survivors who were male or female, not the probability of survival for each group. Since there were many more male passengers overall, this chart hides the critical denominator and can lead viewers to conclude that women dominated survival incorrectly. In the actual event, they did, but you miss out on how many lives were lost as well. The pie also adds further complications to the issue, as it makes comparisons of relative proportions more difficult and provides no context for the total passenger counts or survival probabilities. The visual also uses color contrast and an explosion from the pie chart to reinforce its narrative, though the critical display does not answer the question most viewers assume. It does not help that it does not specify whether this was from the test or training data. Or the whole data. No indication. The use of a pie chart further compounds the issue, as it makes precise comparisons of proportions more difficult and often oversimplifies complex relationships by giving us only a slice of the picture, with no extra data to make comparisons, see trends, or draw our own conclusions. Altogether, this chart demonstrates how even accurate numbers can be framed to shape perception and subtly mislead an audience.

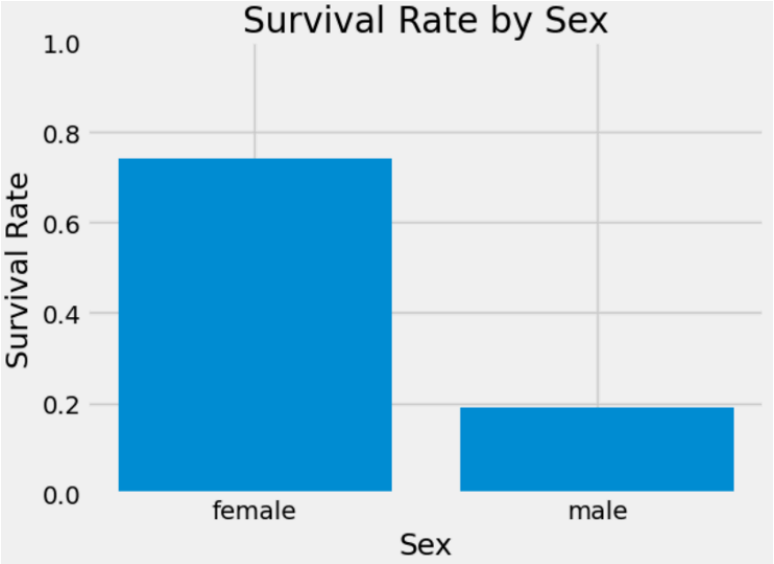


The visualization, "'Huge' Survival Gap by Sex (Truncated Axis)," is a misleading representation of survival rates by gender. At first glance, the graph appears to show an overwhelming difference between female and male survival. The female bar towers over the male bar, visually suggesting an extreme gap in outcomes. However, this dramatic contrast is largely driven by the truncated y-axis, which begins well above zero instead of using the full 0 to 1 range appropriate for survival rates. By cutting off the lower portion of the axis, the visual exaggerates the magnitude of the difference. Although the underlying survival rate calculations are accurate, the scaling choice amplifies the perceived disparity. This design decision manipulates the viewer's perception, making the gap appear larger and more disproportionate than it actually is. The visualization demonstrates how axis truncation, even without altering data values, can distort interpretation and lead audiences to draw stronger conclusions than warranted.

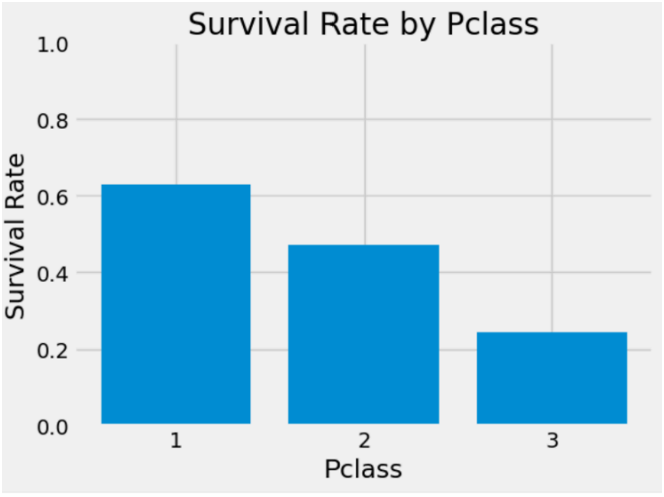


The visualization, "Pclass With 'Most Survivors' (Counts, Not Rate)," is also misleading because it relies solely on raw survivor counts.This bar chart presents the number of survivors within each passenger class and highlights that first class appears to have the highest total number of survivors. While the counts themselves are accurate, the visualization fails to account for the total number of passengers in each class. By using raw counts rather than survival rates, the graph ignores

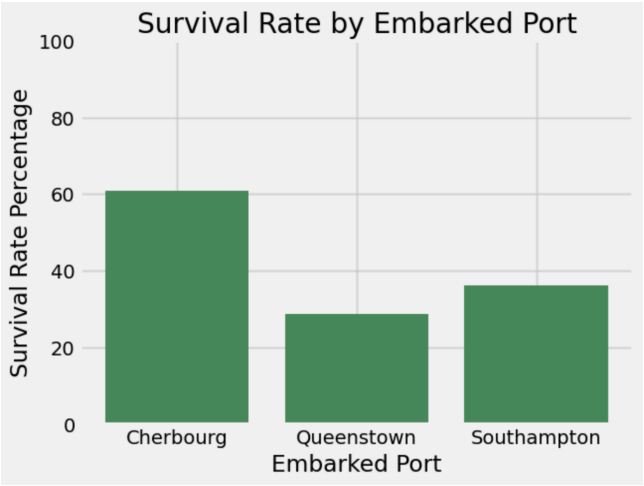
the underlying group sizes, which vary substantially across classes. This omission can lead viewers to incorrectly assume that first class passengers were inherently the most likely to survive. In reality, survival probability must be evaluated relative to the total number of passengers within each class. Without presenting proportional comparisons, the chart oversimplifies the narrative and risks misleading the audience about the true relationship between passenger class and survival.



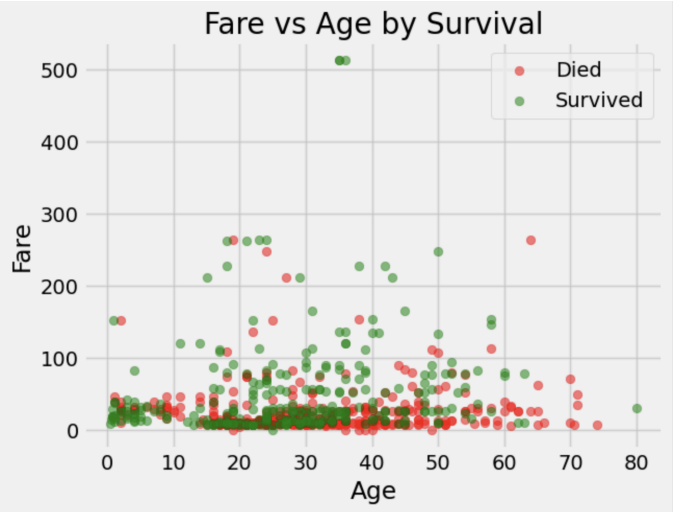
The visualization, “Survival Rate by Sex,” is an accurate and transparent representation of gender-based survival differences. Unlike the truncated-axis graph, this chart uses a full y-axis range from 0 to 1, preserving proportional integrity. By displaying survival as a rate rather than a raw count, the visualization communicates the true probability of survival for each gender. The difference between female and male survival remains substantial, but it is presented without exaggeration. The consistent scaling and straightforward bar design allow viewers to interpret the magnitude of the gap accurately. Because the axis is not manipulated and the data are normalized, this visualization provides a clear and honest depiction of the survival advantage observed among female passengers.



The visualization, “Survival Rate by Pclass,” accurately presents proportional survival outcomes across passenger classes. This chart displays survival rates for first, second, and third class passengers using a full 0 to 1 y-axis scale. By focusing on rates rather than counts, it allows for fair comparison across classes with different population sizes. The visual reveals a clear gradient: first class passengers experienced the highest survival rate, followed by second class, with third class having the lowest. The consistent axis scaling ensures that differences are not overstated or understated. Because the graph accounts for proportional differences and does not hide any portion of the data range, it provides a transparent and analytically sound comparison of how ticket class influenced survival probability.

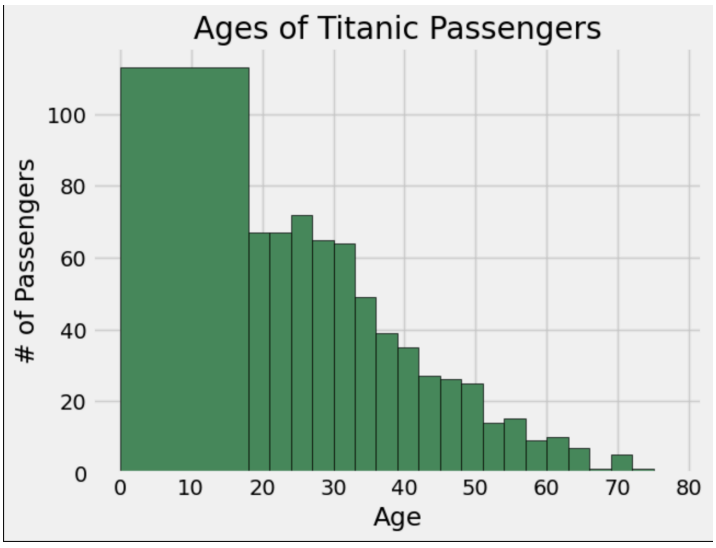


The visualization, “Survival Rate by Embarked Port,” accurately presents the survival rate of Titanic passengers by the port they embarked from. Using the survival rate helps to avoid any skewed data from more passengers coming from one port over the others. The y-axis of this graph shows the survival rate of the passengers as a percentage. There is no modification to the axis, so it can keep in line with the goal of transparency. The color is intentionally the same green, so it can help convey the idea that each bar is representing a survival rate. All ports in the data set are presented, so there is no cherry-picking. I chose a bar chart of a pie chart for this problem because I wanted to convey the individual survival rates as opposed to comparing the percentage of survivors from each port.



The visualization, “Fare vs Age by Survival,” accurately shows the comparison and ticket fare and age of passengers on the Titanic. The survival status of each passenger is encoded by color. Using easily interpretable colors like red for death and green for survival helps to communicate the main idea. The scatterplot above contains every non-null data point available, so no information is hidden. This means there is no cherry-picking to point out certain groups over others. This presentation allows for the data to be presented in a way that does not show any misleading narratives.

The visualization, “Survival by Sex”, shows the survival counts of the passengers aboard the Titanic by sex. This graph is intentionally misleading because it used the color red to represent passengers that survived, and the color green is used to represent passengers that died. In the paper “Principles of Effective Data Visualization,” the author dedicates a principle to the importance of color in graphs. Someone that scans this graph quickly may assume that green means survived as it is associated with positive outcomes.



The visualization, “Ages of Titanic Passengers,” displays the counts of Titanic passengers by age in the form of a histogram. This graph uses the manipulation of bin size to exaggerate the perception of the number of child passengers aboard the Titanic when it crashed. In the article “The 5 Most Important Principles of Data Visualization,” the author describes the importance of telling the truth by not going against basic conventions. This graph intentionally goes against the convention of making each bin the same size. This distorts the truth by making it appear that there were significantly more child passengers aboard the Titanic when it crashed.

In conclusion, assumptions should not be made as they mislead the ultimate goal of those creating visualizations. If a movie tells a story of a historic event, use the data to answer questions with accurate facts. Deceiving with visualization not only tricks the audience but also rewrites history. It is absolutely necessary to use data ethically because it can create real world consequences that put real people at risk. Using misleading visualizations can create a narrative that puts some people at unwarranted disadvantages as well. As data scientists, it is a crucial part of our job to make sure visualizations ensure that all people can be represented and nobody is misinformed. In the end, the Titanic dataset shows data does not lie. Rather, instead of misleading an audience, ethical visualizations use that data to show precise context, accurate scales, and enough information for valuable insight into that historical ship wreck.