Exploring Bias in Datasets

Group 5

Introduction:

Data Visualization plays a critical role in how insights are communicated and understood when presenting data findings. Good visualizations can be a powerful tool for communicating information because they transform raw numbers into patterns that viewers can interpret quickly. Visual representations of data shape how audiences understand trends, compare groups, and draw conclusions. However, the same design choices that make visualizations compelling can introduce bias, distortion, or misunderstanding if not constructed carefully. Things such as scale, color, grouping, and framing all can influence perception which can be used to manipulate an audience into a misleading interpretation. Using the Titanic passenger dataset as a case study, we explore the effects of both accurate and misleading visualization techniques to demonstrate how different design decisions impact interpretation by the audience. By comparing clear and transparent visuals with intentionally distorted ones, the analysis highlights the ethical responsibility that comes with presenting data and emphasizes the importance of accuracy, clarity, and thoughtful design choices in communicating quantitative information.

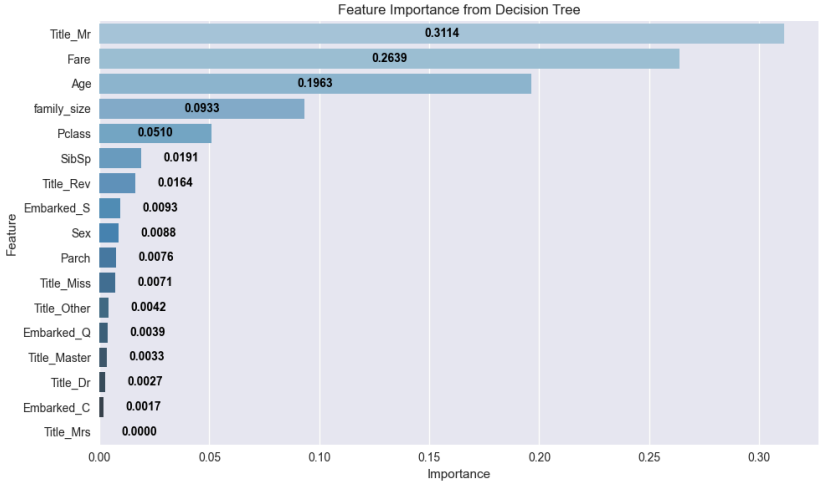
Summary of the Dataset:

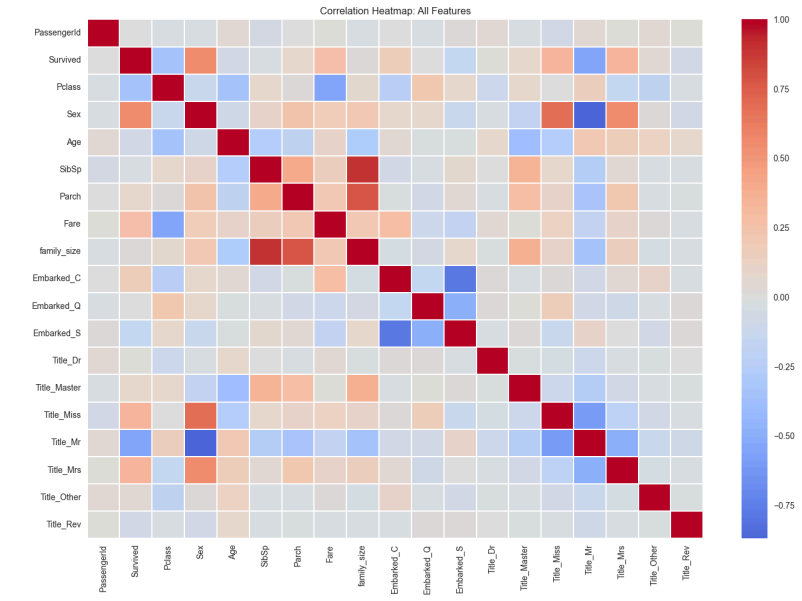
For this visualization project we chose the Titanic dataset. The Titanic training dataset contains 891 passenger records and 12 variables used to predict survival during the Titanic disaster. Each observation represents a single passenger and includes demographic information, socioeconomic indicators, family structure, and travel details. Some of the features in the set are age, sex, number of children, ticket class, and cabin. The target variable for this set is Survived. Survived is a binary variable with approximately 38% of passengers surviving and 62% not surviving which creates a moderately imbalanced classification problem. When looking at the raw data we see the dataset is missing about 77% values for cabin and about 20% in age, requiring some preprocessing before modeling. Title was engineered in the dataset by extracting the title from the name column with some cleaning incorporated. In the fare column, some passengers in third class had missing values for the feature. So the missing values were filled with the median fare for that class. Missing age values were handled by creating a dictionary of SibSp and Parch pairings as keys and the median age of that pairing as values. If an entry had a SibSp and Parch pairing that existed in the dictionary, then the age was filled with the median age for that pairing. SibSp is a count of the number of siblings and spouse that was aboard the Titanic with the passenger and Parch is a count of parents and children. Family size was calculated by adding the number in SibSp and Parch and adding 1 to account for the passenger themself. Overall, the dataset combines mixed datatypes, missing or null data, and meaningful socioeconomic and demographic information, making it a well suited dataset for classification modeling, feature engineering, and exploratory analysis to understand the factors that influenced survival. For modeling, we used the decision tree classifier to help make and see how predictions would do and to see what features are important to the model and gain insights from the correlations.The decision tree model needed to be processed for modeling by using all the cleaned data sets and the gender_submission data (merged into the test dataset) by using

feature engineering and encoding to prepare for modeling. The feature engineering used was to make dummy variables for the Titles and Embark features where it would be helpful in finding out which one makes a big importance or correlation with other features. The encoding was to make the Sex feature from being text to numerical and so male became 0 and female became 1. The y variable for the decision tree was Survived feature see if the model can predict the outcome based on all the other features. This model led to find out feature importance and what features correlated with each other to gain additional insight into who survived that night during the sinking of the Titanic.

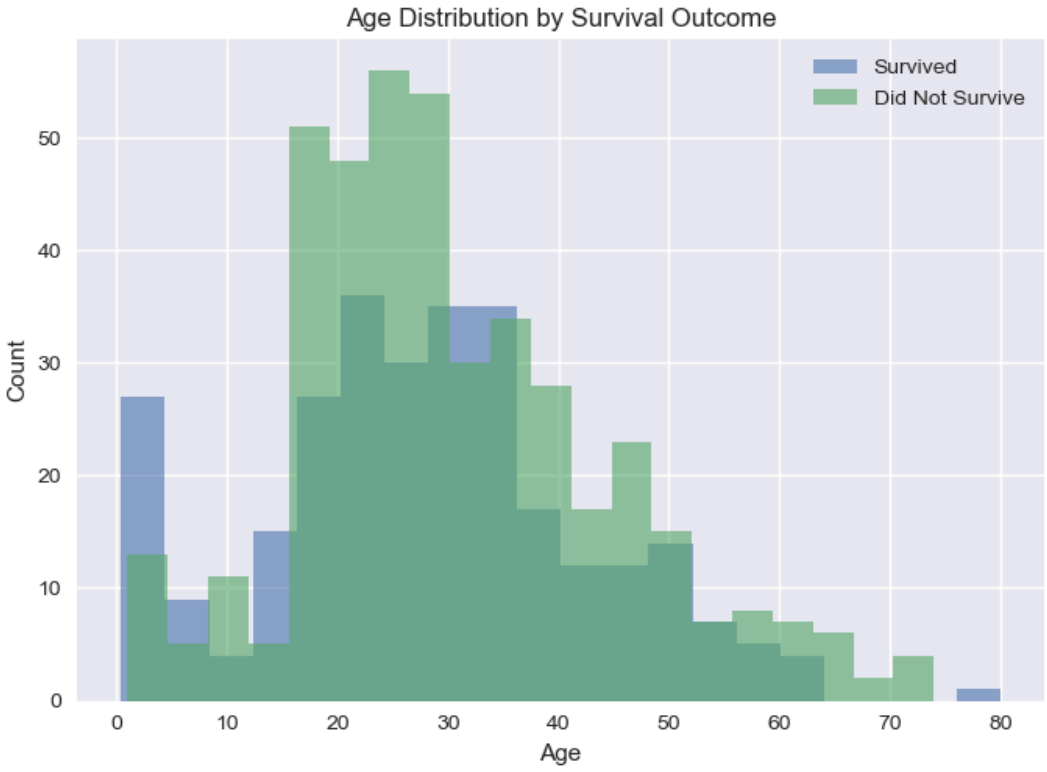
Description of the visualizations and their implications:

Accurate:





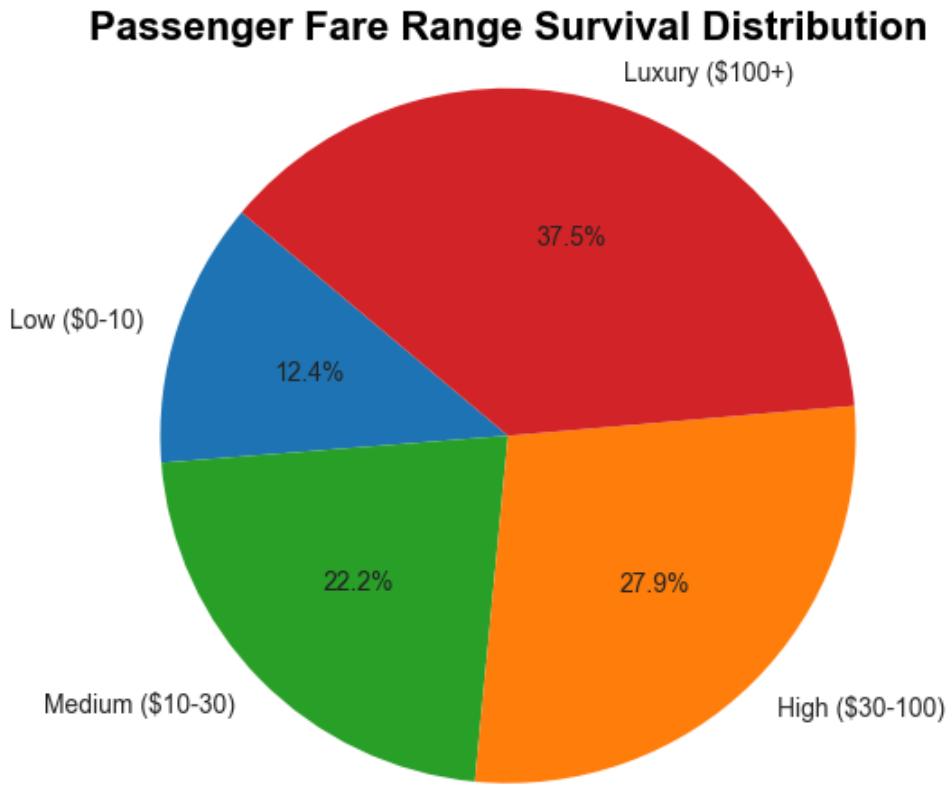
For the first two accurate visualizations, Feature Importance from the Decision Tree and Correlation of All Features, a decision tree classification model as built to evaluate which variables most strongly influenced survival predictions, while a separate correlation matrix was used to examine linear relationships among the features. The first visualization is the feature Importance (for this model we had to make dummy variables for embarkation and title) and the top three feature importances were Title_Mr, Fare, and Age had the biggest influence on the survivability in the model. This is pretty accurate in that Title_Mr is indicating that this would be the top feature with men having the lowest survival rate compared to women and children. This leads to age on the ship if you were in your 20s to 30s you had a lower survival rate compared to children and older. Then for fare this has to do with how unfair the lifeboat loading was where the top rich had a high survivability compared to lower class who paid a lower fare. This all correlated with each other shown in the correlation matrix visualization where it is shown that sex and titles are correlated with each other having Title_Mr. having the lowest negative correlation to survivability with the sex feature. There is a high correlation that shows children had the highest survival rate out of anyone with familysize and sibsp(siblings/spouses) with parch(parents/children) had high correlation with each other in terms of survival. These visualizations are accurate for the technical audience and background knowledge audiences but not for the public which would have a hard time trying to figure out what the visualizations mean.



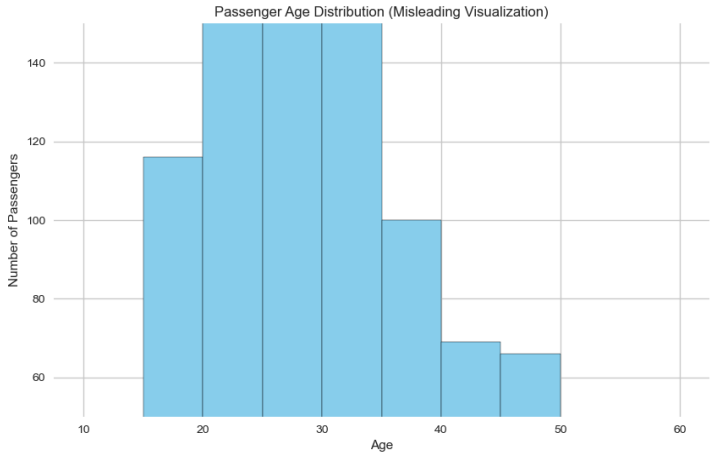
This overlaid histogram of age by survival outcome is an effective positive visualization. It is an effective visualization because it allows for direct, intuitive comparison between two groups while maintaining clarity and minimizing distortion. By using consistent bin widths, a shared axis, and transparency the chart makes it easy to observe both the overlap and the difference in age

distributions without misleading the viewer. The viewer can also quickly look at this and see that survivors skew slightly to the right showing that younger passengers survived more often while also showing that non-survivors are more concentrated in adult age ranges. This visual doesn't hide the large overlap supporting a more complicated interpretation instead of an oversimplified model. This design supports truthful interpretation rather than exaggerating contrasts which aligns with Swadia's emphasis on clarity, appropriate scaling, and truthful representation in effective data visualizations (Swadia, n.d.). The visualization also avoids common misrepresentation issues such as misleading axis or scaling, unnecessary embellishment, and clutter that can negatively affect insights as discussed by Treesak (n.d.). Having both groups on the same frame of reference allows the data to speak for itself. The visualization communicates insight efficiency, supports honest interpretation, and demonstrates best practice in transparent comparative analysis making this a solid and effective visualization for this dataset.

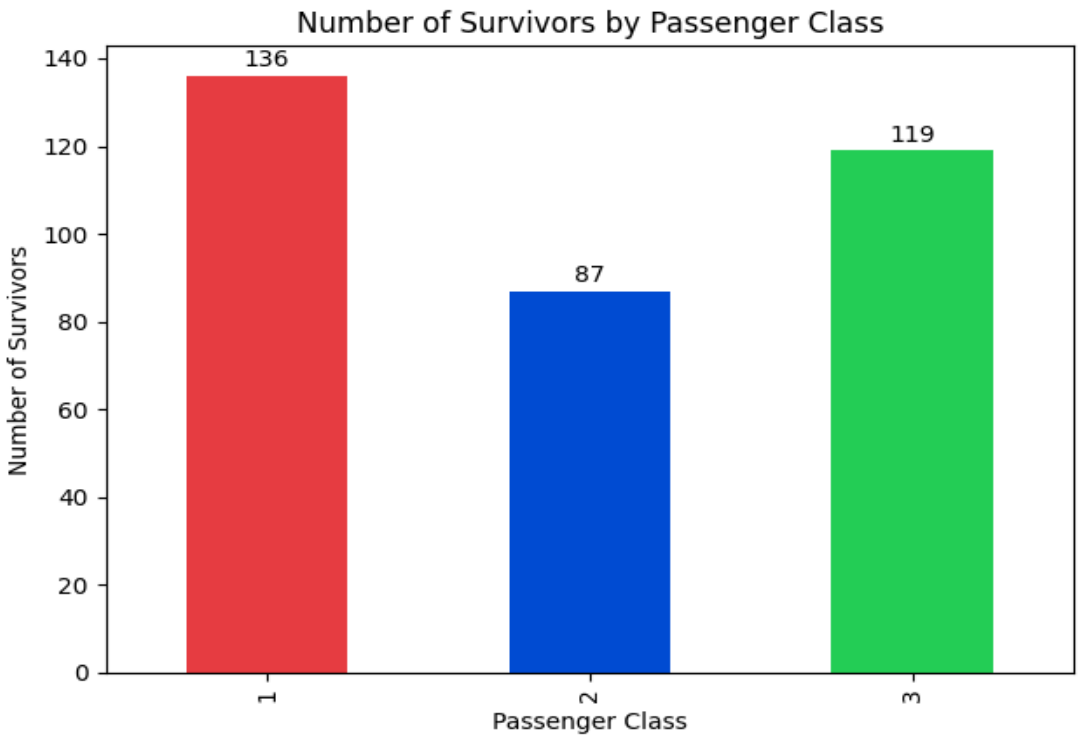
Misleading:



This visualization of ‘Passenger Fare Range Survival Distribution’ is very misleading to the audience that would be reading it. This is showing the more you paid for fare the higher the chance of survival during the sinking, also saying there were a lot more higher fare passengers compared to the lower classes that survived. What makes this a misleading visualization is one it is a pie chart and two how the data is shown in the pie chart. There is also hidden misleading where this pie chart had averaged the fares and binned the fares into four categories. This is already misleading with averaging the lower class where they had the most passengers compared to the higher classes. Pie chart is not a favored visualization due to the user could be confused since portions of the chart are similar to one another and text modifications make it hard to read what the actual data is saying (Treesak, Y (n.d.)).



This visualization is misleading for two primary reasons: baseline shifting and cherry picking. First, the truncated y-axis, which does not start at zero, exaggerates differences between age groups by making relatively small variations in passenger counts appear much larger than they actually are. In count-based visualizations such as histograms, failing to begin the y-axis at zero distorts magnitude perception and can lead viewers to incorrectly conclude that certain age groups are disproportionately represented. Second, the data appears to be filtered which cherry-picks a limited age range (approximately 15–50), which excludes younger and older passengers. This selective filtering alters the overall shape of the distribution and can create a misleading narrative about the demographic makeup of passengers. Together, these design choices shift the visualization from objective analysis toward persuasive storytelling. Ethically, such practices can mislead stakeholders, reinforce confirmation bias, and erode trust in the analysis. Responsible data visualization requires transparency about filtering decisions, accurate scaling, and faithful representation of the full dataset to ensure that conclusions are based on honest and reproducible evidence.



Raw Counts Instead of Rates

This chart displays the total number of survivors per class rather than the survival rate. At first glance, the numbers appear relatively close. First class has 136 survivors and third class has 119. A viewer could reasonably conclude that survival outcomes were somewhat comparable across classes. However, this interpretation ignores the population size. Third class had 491 passengers, while first class had only 216. First class passengers were more than twice as likely to survive. This connects to a common visualization mistake discussed by Treesak (n.d.), on data visualization pitfalls; comparing totals across unequal groups can distort interpretation because the denominator is missing. Without context, viewers cannot properly evaluate fairness or inequality because the design hides the structural disadvantage faced by 3rd class passengers.

The second distortion comes from color choice. Red is used for the highest survivor count, while green is used for the lower one. In many cultural contexts, people associate green with good and red with bad. This relates to a basic visualization principle: design elements should support interpretation, not complicate it. When color encoding does not align with known expectations, viewers might subconsciously interpret the colors the wrong way, especially if they are just scanning the graph quickly.

Ethical considerations in visualization design:

When designing visuals, it is important to think about ethical considerations and implications that design choices may have on the data and analysis being presented. Visualizations should be appealing and attractive to the eye, but it is important that the visualization should never take away from conveying the original message meant to be conveyed by the data. It is best practice to use an appropriate color scale with high contrast. For visuals that are colorized with several shades of a single color, a sequential color palette will work best and darker colors indicate greater emphasis. For visuals that represent categorical data, a qualitative color palette can use two distinct colors and utilize white as a middle color between the two. We also learned that visuals can and should include good captions that reference only the visual they are attached to. Avoid practices that may lead to persuasion of audiences or coercing them into blindly accepting your claims such as cherry picking and baseline manipulation. By following these ethical practices, any data visualization has the capability of being accurate and ethically correct.

References (APA Style) (used purdue owl to help)

Swadia, S. (n.d.). The 5 most important principles of data visualization.

Treesak, Y. (n.d.). 10 common data visualization pitfalls to avoid.