Dave ON DATA

Data Visualizations for Machine Learning

Greetings!

The power of data visualization cannot be overestimated. The human brain is optimized for visual pattern recognition – which is why “a picture is worth 1,000 words.”

This power extends beyond executive dashboards into the world of exploratory data analysis (EDA). Tools like Tableau, Power BI, and the mighty ggplot2 in R (my favorite) make creating powerful EDA visualizations quick and easy.

In my experience, this is especially true when you are using machine learning (ML) algorithms like decision trees and random forests to analyze your business.

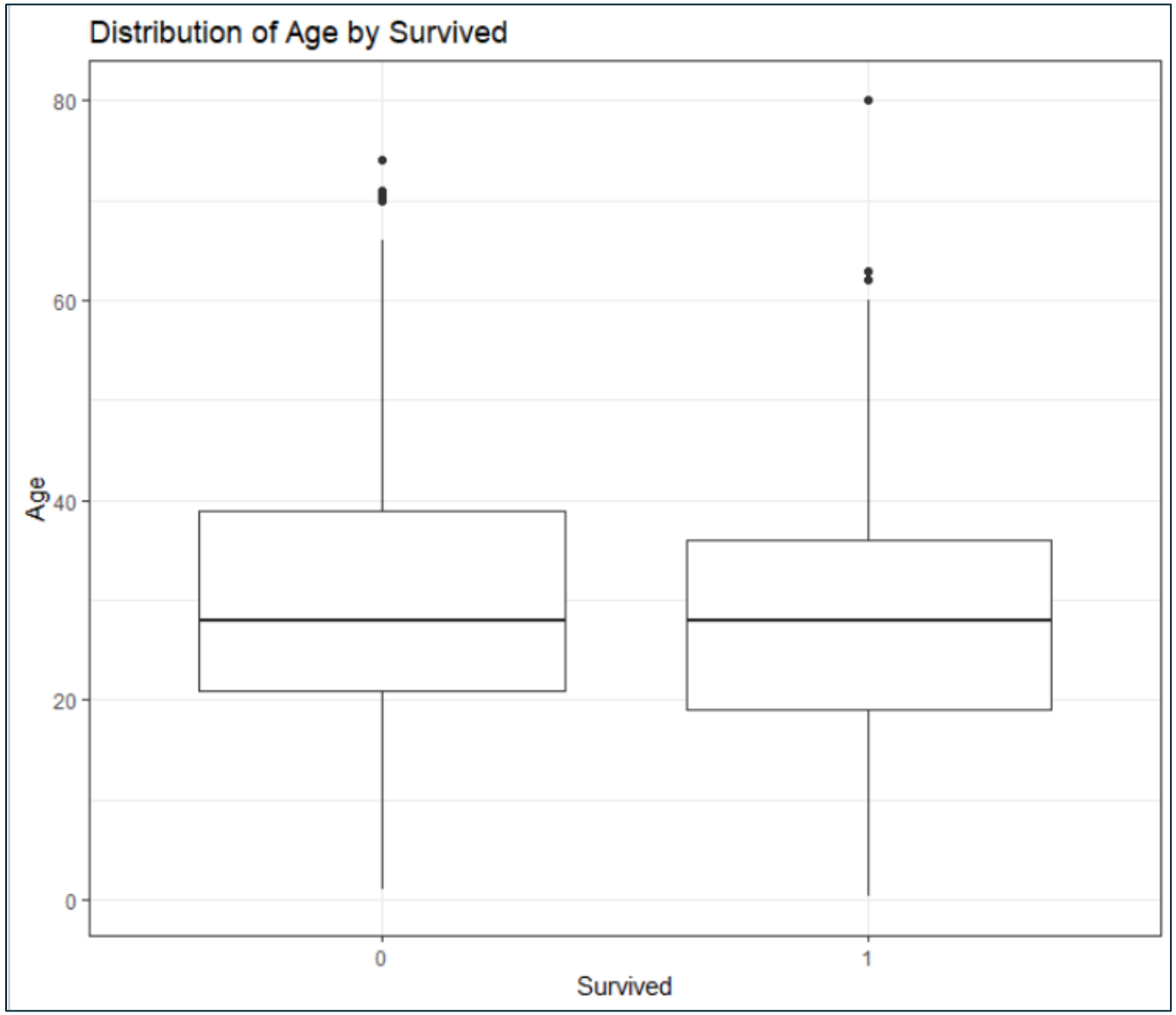
This deck illustrates the visualizations I teach in my “From Excel to Machine Learning Bundle” using R’s incredibly powerful ggplot2 library with Kaggle’s famous Titanic data set:

- https://bit.ly/ExcelToMLBundle

Stay healthy and happy data sleuthing!

-Dave

Box Plots

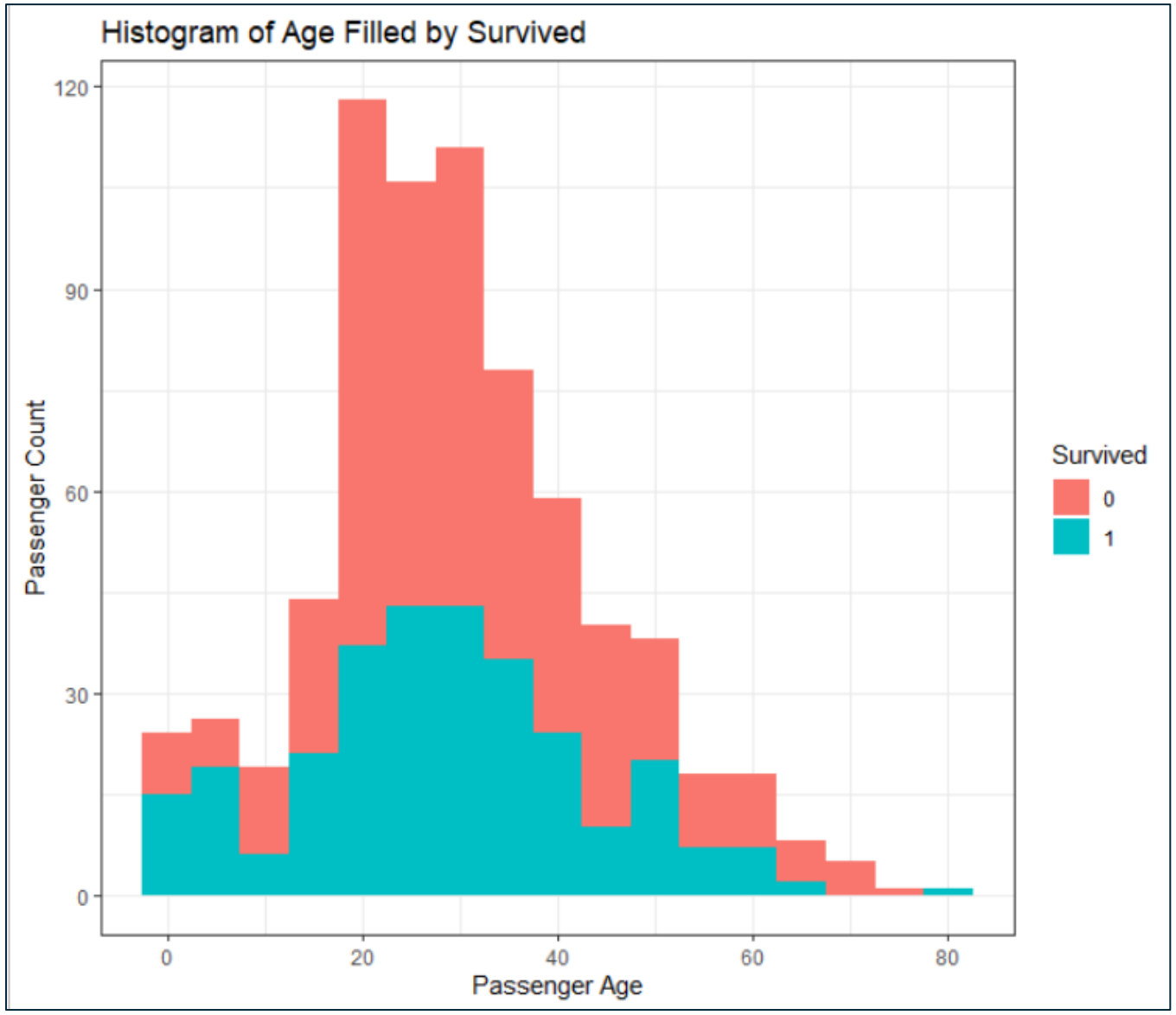


While not the ideal data visualization in every scenario (e.g., drug trials), the box plot remains a very handy tools in business analytics.

Box plots are a good example of a 2-dimensional data visualization. In this case, *Survived* on the X axis and *Age* on the Y axis.

This visualization illustrates that the overall distribution of the Age feature is quite similar between passengers that survived and those that did not.

Filled Histograms



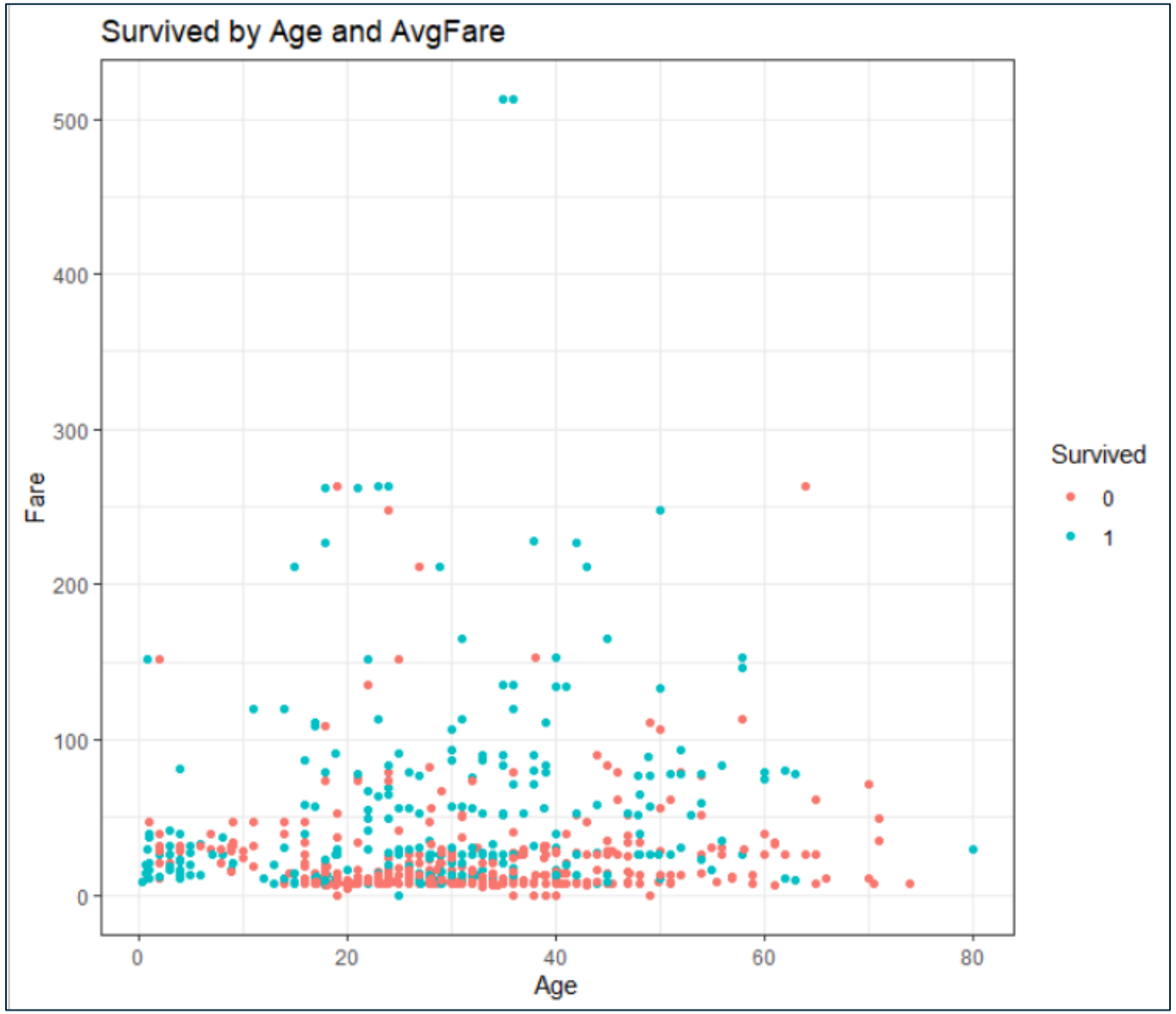
Filled histograms communicate a ton of information quickly.

Filled histograms are another example of a 2-dimensional data visualization.

A pattern in using data visualization with machine learning is that more dimensions in your visualizations offer more power/insights.

This visualization illustrates, in general, that the youngest passengers had better survival rates.

Color-coded Scatter plots



Color-coded scatter plots are and example of a 3-dimensional data visualization.

In this case – *Age*, *Fare*, and *Survived*.

Arguably, this visualization doesn't provide a ton of insight.

There's a reason why they call it EDA and not, "guaranteed data insights."

Proportion Bar Charts

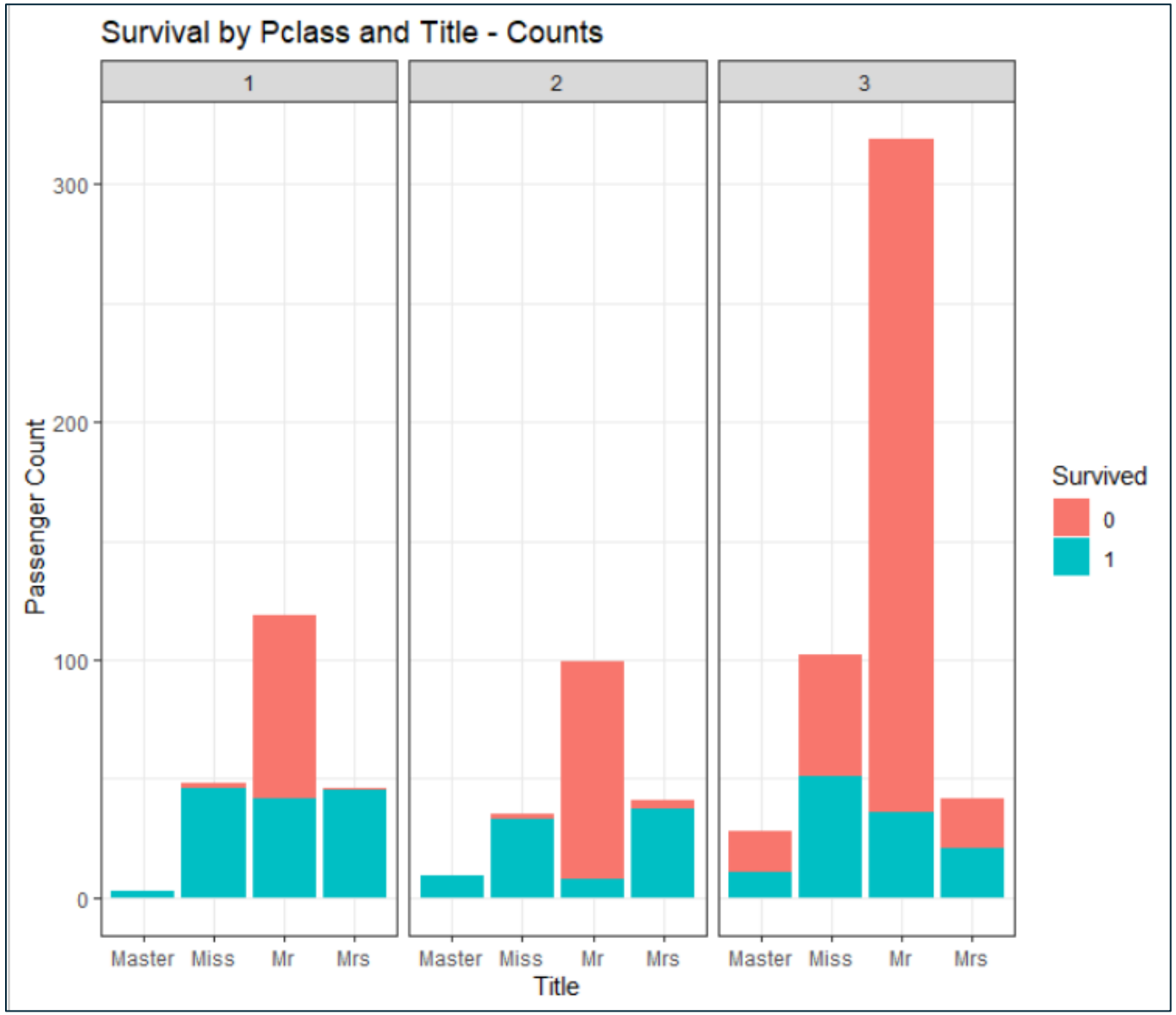
Proportion bar charts are a powerful data visualization.

When these visualizations become multidimensional, their power only grows.

This example visualizes three categorical features – *Pclass*, *Title*, and *Survived*.

This plot clearly illustrates that the combination of *Pclass* and *Title* has a profound effect on passenger survival.

Count Bar Charts



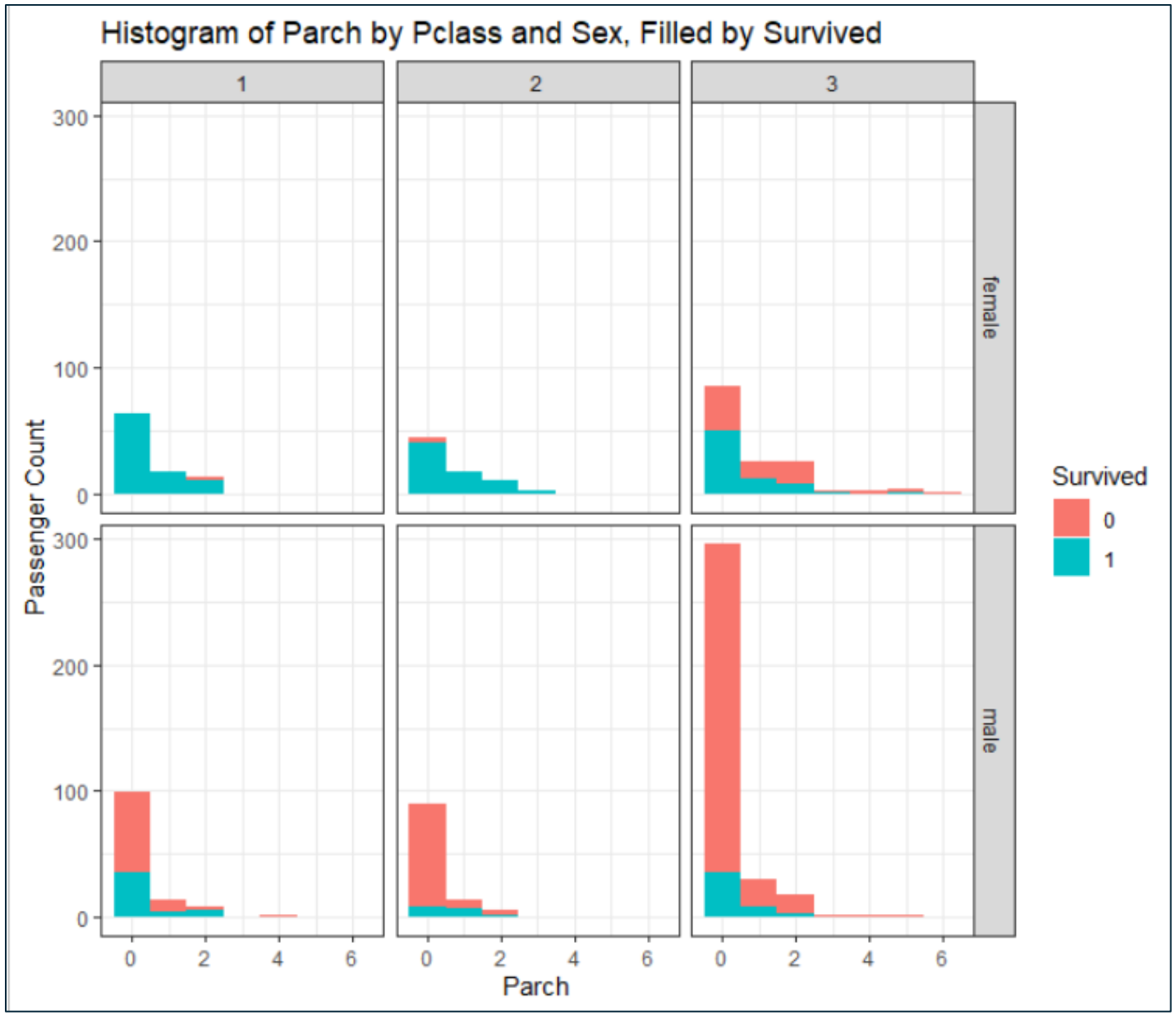
When analyzing data with machine learning, you can't rely on proportion bar charts alone.

You also need to understand the counts of observations to gain a sense of "the gravity of the data."

For example, tree-based ML algorithms consider observation counts as they learn from the data.

This plot clearly illustrates that the overall gravity of the Titanic data are passengers with the title of "Mr."

Getting Creative



As mentioned previously, adding more dimensions can often lead to new insights when analyzing data.

This is an example of a 4-dimensional data visualization of *Parch*, *Pclass*, *Sex*, and *Survived*.

The visualizations tells multiple stories simultaneously:

1. Survival tends to be associated with smaller *Parch* values.
2. Females in 1st & 2nd class overwhelmingly survive regardless of *Parch*.
3. Most male passengers have *Parch* of 0.