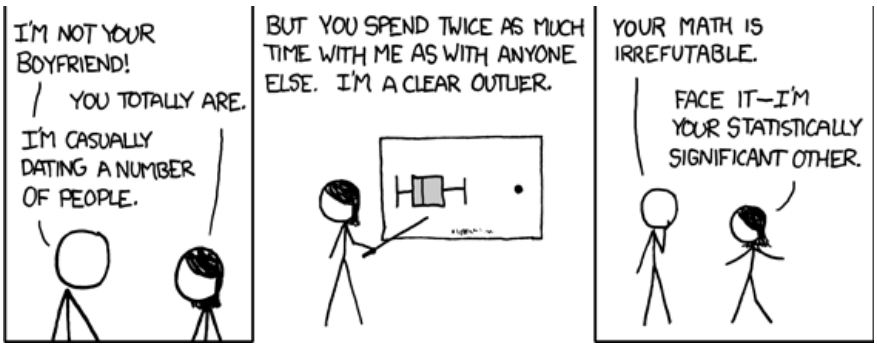
Marcon Laforet Follow
A programmer with a passion for all things data science.
Dec 24 · 5 min read

How to visualize distributions

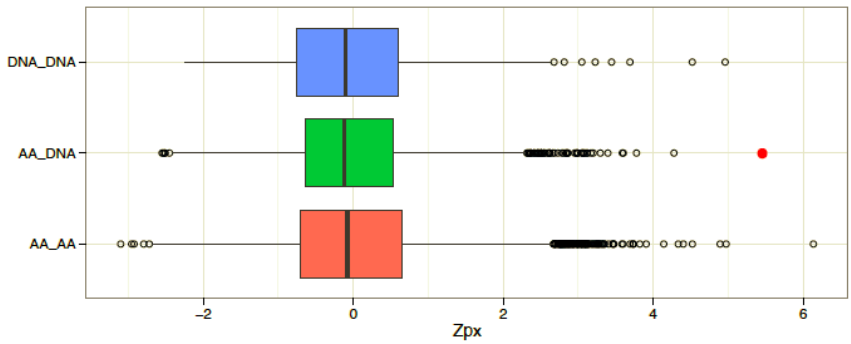


https://i.stack.imgur.com/3ngU8.png

You have munged all the necessary data into a nice clean format, you've appropriately performed a snazzy statistical analysis and now it's time to analyze the results. This is where visualizing your data comes in handy. Informative data visualizations not only reveals novel insights, maybe you were dating someone and didn't even know it, but they are truly invaluable when it comes time to communicate your findings to your boss/client.

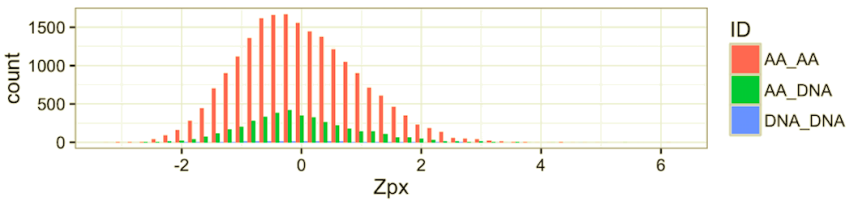
This post will specifically look into a visualization task that I've faced time and time again. Over the years of analyzing data, I often find myself wanting to compare and contrast multiple distributions of numeric data. This can be tricky depending on how different the distributions are, compounded by the vast number of possible representations methods for distributions (http://www.darkhorseanalytics.com/blog/visualizing-distributions-3). I commonly have one of two objectives when comparing distributions, either I want to highlight differences in their outliers or, often subtle, differences in their respective spreads. Maybe I want to show how datasets gathered with distinct criteria responded differently to a statistical procedure or how applying a statistical correction improved a scoring function.

I tend to favour box plots if I'm interested in comparing outliers. Box plots show the overall spread of the data while plotting a data point for outliers. This physical point allows their specific values to be easily identified and compared among samples.

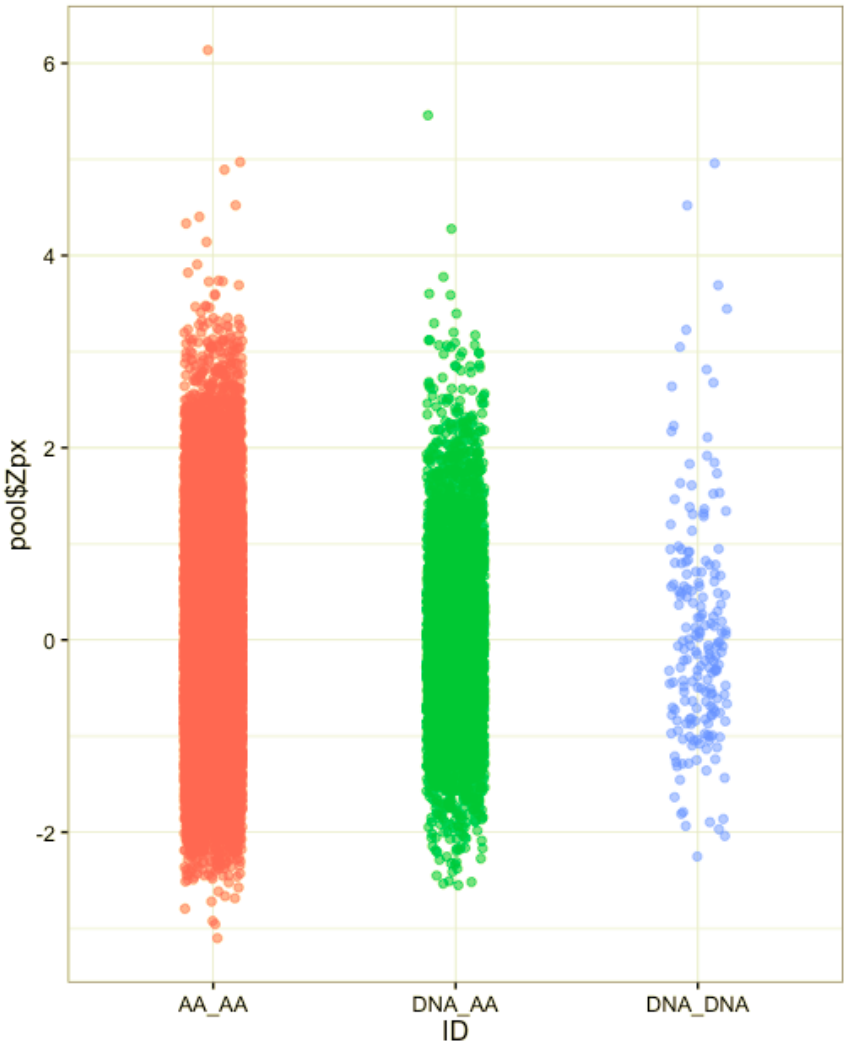


Let's ignore what the data actually means as that's not important. You can see that the spread of the distributions are more or less equal and that the outliers are easily compared. The distribution coloured red/magenta has the most extreme outlier followed by the point that I coloured red from the green distribution. For this analysis, the red distribution had been previously calculated and I was able to reproduce their data by observing the extreme outlier. The red point however, was a novel observation.

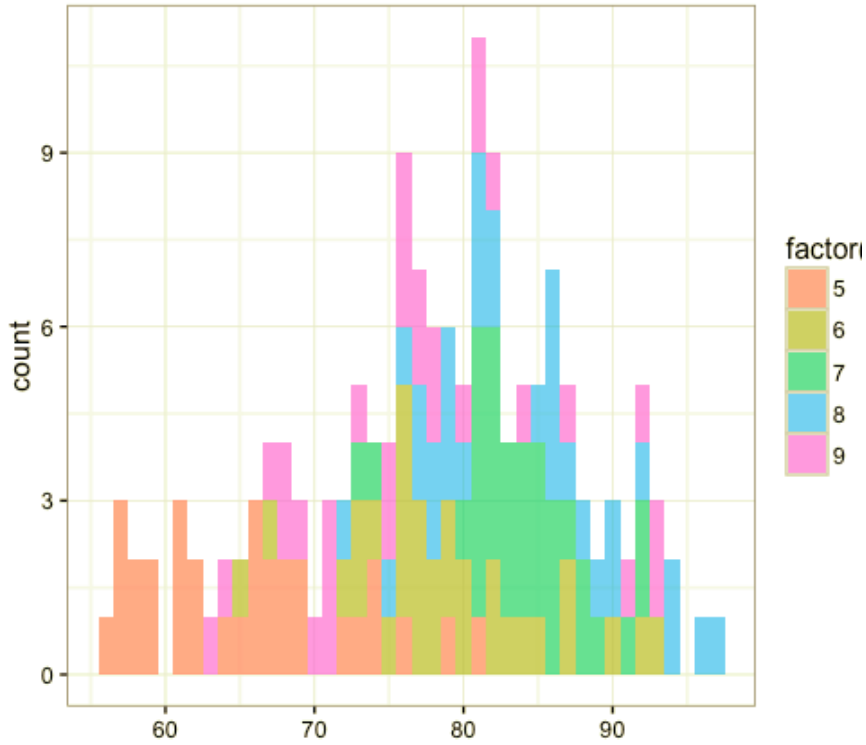
If you're a person that pays attention to the plot axis and understands a little stats, then you might have realized that I applied a statistical transformation to my dataset in order to amplify the differences in the distribution outliers. I transformed my numerical distribution to a z-score. A z-score transforms the data points by measuring the number of standard deviations away they are from the sample mean.



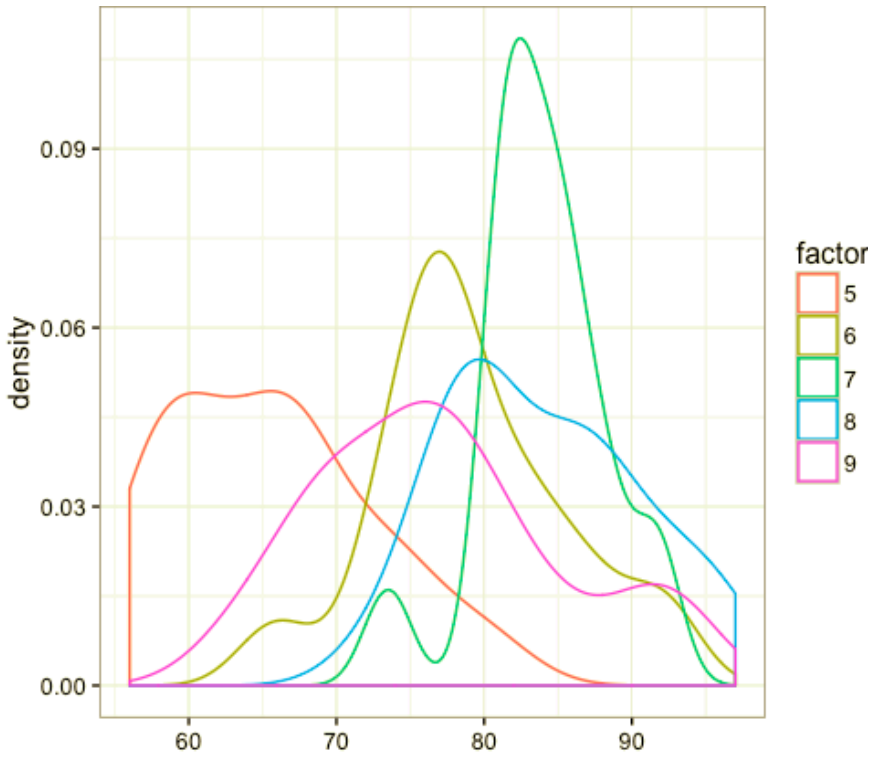
The first visualization I usually make for distributions is a histogram. You can see here that this is a terrible and uninformative way to look at the data. The differences in the sample sizes between the different groups makes them incomparable using this method. It is so extreme that you can no longer see the blue distribution. This visualization also fails at comparing or even seeing the outliers. The only thing I can conclude from this visual is that the red and green distributions have roughly the same mean.



Although I think that box plots were the best option in this case, they can seem very formal and people often don't know how to interpret them properly (Interquartile ranges, distributions, say what?). Additionally, box plots give no insight into the sample size used to create them. A strip plot can be more intuitive for a less statistically minded audience because they can see all the data points. This plot also gives an insight into the sample size of the distribution. I like to apply jitter and opacity to the points to make these plots more appealing.

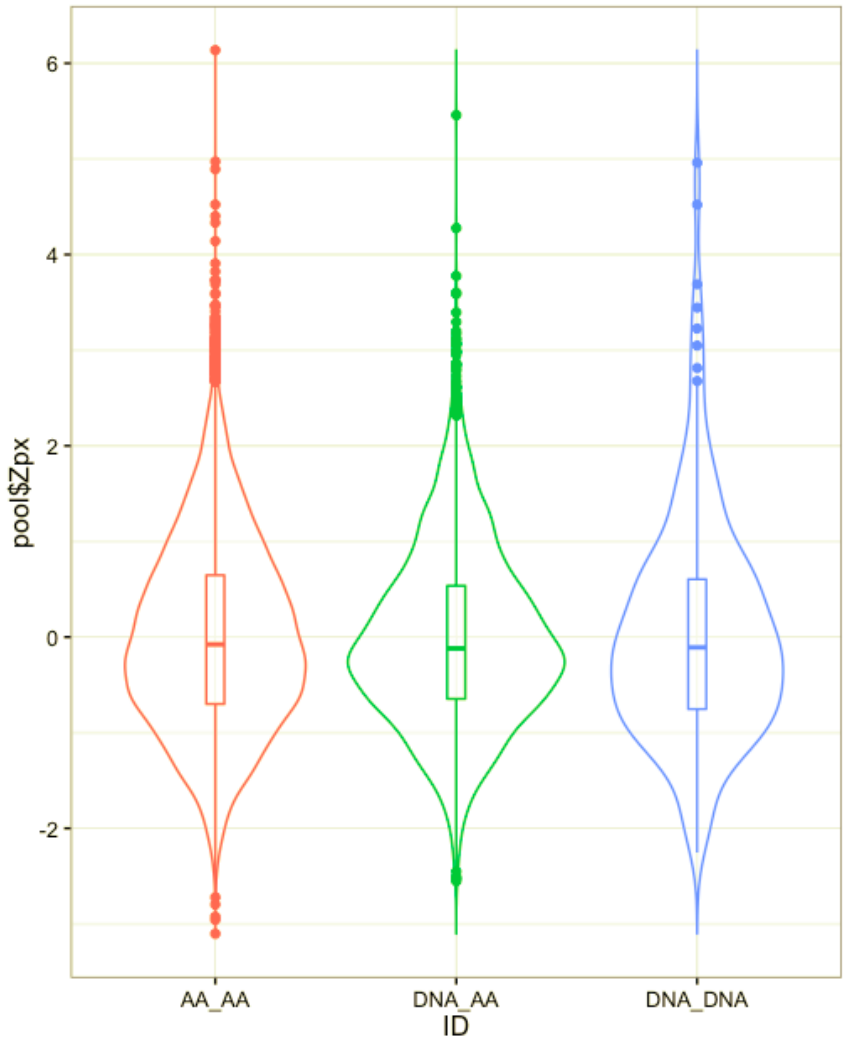


Let's jump into our second case, where we are interested in comparing the spreads of the distributions. Here, histograms are a good option if the distributions being compared have the same sample size and you are making at most 3 comparisons. Otherwise you will end up with a really busy plot that makes it very hard to see the data.



I gravitate towards kernel density plots with no fill for these cases. It isn't that pretty but you can actually see and compare the distributions. To overcome this in a recent project, I decided to implement a spin on the histogram and use a variation called the step plot that worked out great. I would suggest changing the way that you are representing your data if your plots are getting unwieldily.

But what if you want it all?! In this case I like to use a violin plot. I have seen these plots becoming more popular and there are many variations that make them even more powerful. They are essentially boxplots that have a rotated kernel density plot around them. Here I plotted the boxplots inside the rotated kernel density plot.



That is all I got for you for now. I made all the plots above using the ggplot2 package in R. I also make quite a few plots in python using matplotlib and sometimes seaborne. The dataset used in case 2 was done using the airquality dataset shipped with R and the other dataset was built by myself for my masters thesis.