We were recently working on a supervised learning problem (i.e. building a model using some features to predict some response variable) with a fairly large dataset. We used base R’s plot and hist functions for exploratory data analysis and all looked well. However, when I started building my models, We began to run into errors. For example, when trying to fit the lasso using the glmnet package, I encountered this error:



We thought this error message was rather cryptic. However, after some debugging, We realized the error was exactly what it said it was: there were NA/NaN/Inf values in my data matrix! ***The problem was that I had expected these problematic values to have been flagged during my exploratory data analysis.*** However, R’s plot and hist functions **silently** drop these values before giving a plot.

Here’s some code to demonstrate the issue. Let’s create some fake data with NA/NaN/Inf values:

n <- 50 # no. of observations

p <- 2 # no. of features

# create fake data matrix

set.seed(1)

x <- matrix(rnorm(n \* p), nrow = n)

# make some entries invalid

x[1:3, 1] <- NA

x[4:5, 2] <- Inf head(x) #> [,1] [,2]

#> [1,] NA 0.3981059

#> [2,] NA -0.6120264

#> [3,] NA 0.3411197

#> [4,] 1.5952808 Inf

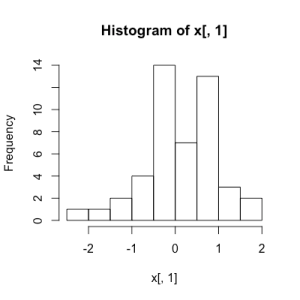
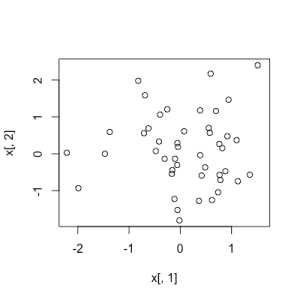
#> [5,] 0.3295078 Inf

#> [6,] -0.8204684 1.9803999

The two lines of code give plots in return, **without** any warning message to the console that data points have been dropped:

plot(x[, 1], x[, 2])

hist(x[,1])

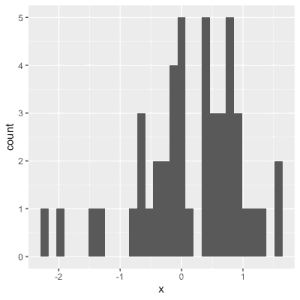


The ggplot2 package does a better job of handling such values. While it also makes the plot, it sends a warning to the console that some values have been dropped in the process:

library(ggplot2)

df <- data.frame(x = x[,1])

ggplot(df, aes(x)) + geom\_histogram()


**Moral(s) of the story:**

1. Don’t assume that your data is free of NA/NaN/Inf values. Check!
2. Base R’s hist and plot functions do not warn about invalid values being removed. Either follow the advice in the previous point or use code that flags such removals (e.g. ggplot2).