# Foundations of Data Analysis

*The University of Texas at Austin*

*R Tutorials: Week 3*

## Scatterplots

In this R tutorial, we'll learn how to make a scatter plot in R, which will show the relationship between two numeric variables. And as a demonstration, we're going to use some variables from our bull riding data set. If you're using a server version of R, make sure it shows up in your files window, down here. But since mine's already saved on my desktop, I'm just going to go into my environment or workspace, say Import data set from a text file, find that CSV file, bullriders.csv, hit Open. And then to shorten the name up since I'm going to be typing it a lot, I'm just going to call this bull. I do have variable names in my first row, so I want to change the heading option to yes. And then I go ahead and hit Import.

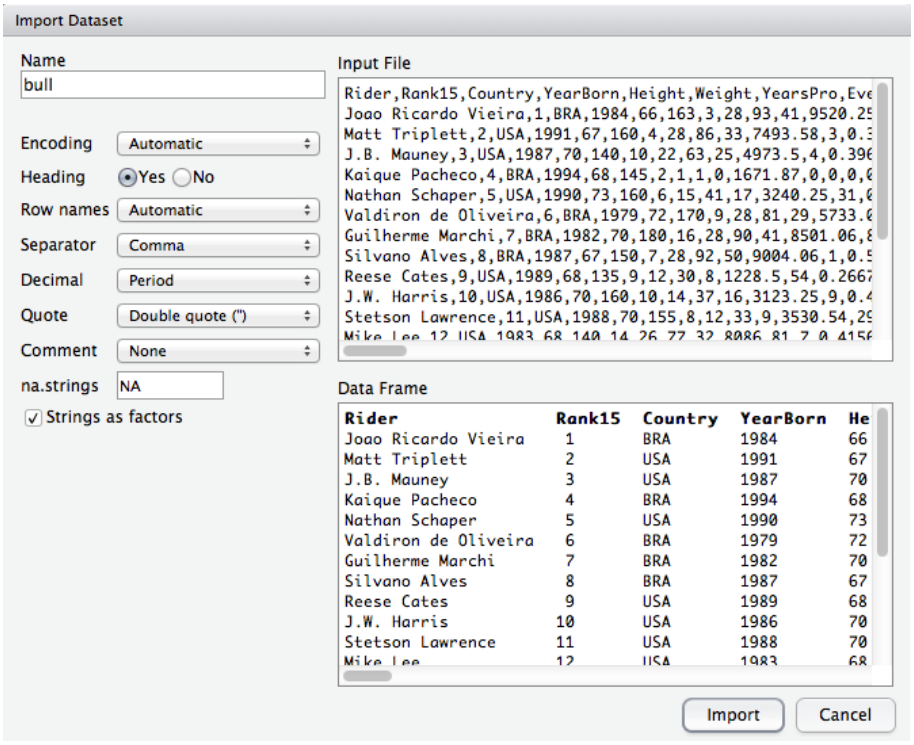



Figure 1: Importing Data in RStudio.

Alternatively, we can use the "**read.csv()**" function from R to bring in the dataset from a local location:

```
bull <- read.csv("~/Desktop/SDSFoundations/BullRiders.csv")
```

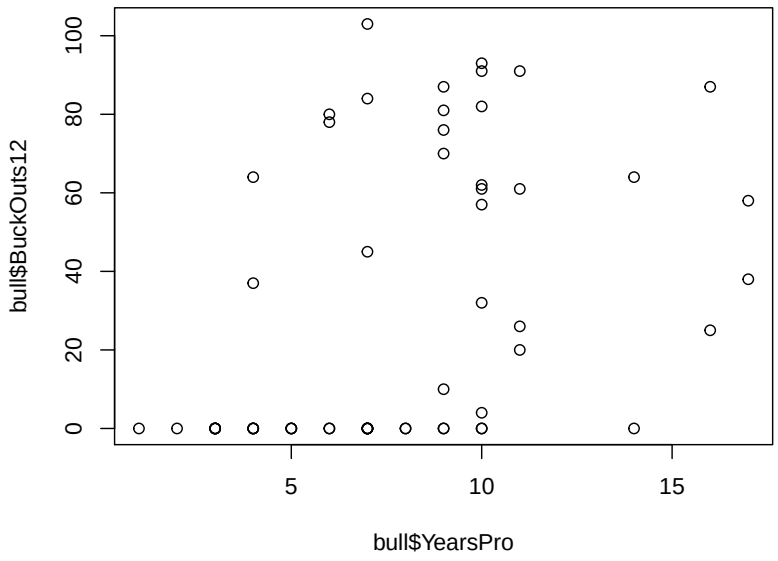
So we see that this data set has 58 professional bull riders. And we've got 44 variables that describe these riders, such as their height and weight, how many years they've been a professional bull rider, how many events they've ridden in for a particular year, how many buck outs, or attempts that they've gone out the gate on the bull for that year, how many successful rides they've had – meaning they've stayed on for 8 seconds – and various other numeric variables that we can analyze here. So for this video, we're going to look

at that buck outs variable. Again, you can think of this as the number of attempts they've tried to stay on the bull for this season. And we want to see if maybe it's linearly related to the number of years they've been professional bull riders, or maybe the number of events they've ridden in that year.

So we're going to look at the relationship between buck outs and those two variables. Before you run any correlation analysis, you want to first confirm that the two variables you're going to analyze don't have some non-linear relationship. Because correlation measures the strength and direction of the linear relationship, you need to confirm that the two variables are indeed linearly related. And the best way to check for the assumption of linearity is to just make a scatter plot of the two variables. So in R, we can use the "**plot()**" function. And if we give it two numeric variables, it will produce for us a nice scatterplot.

So let's start with our YearsPro variable. So the first thing we want to give the plot function is our first variable. That will go on the horizontal or x-axis. It'll kind of act as the independent variable. And then we're going to use a comma to separate out the second variable, which will be BuckOuts12 (number of buck outs for 2012 season). And this variable will be on the vertical axis, or the y-axis, and act as the dependent variable in our graph. Now just like hist function another plot functions in R, we can give axis labels and a main title to the graph. And we're going to separate all these options out with commas. And these options are called by the same things they were called with the histogram function. "xlab=" will give us an opportunity to label our x-axis with the variable name. "ylab=", can do the same for the y-axis. And then, finally, "main=" is where we can give it an overall title for the graph. Now, make sure you have a closing parentheses on this function.

```
plot(bull$YearsPro, bull$BuckOuts12)
```

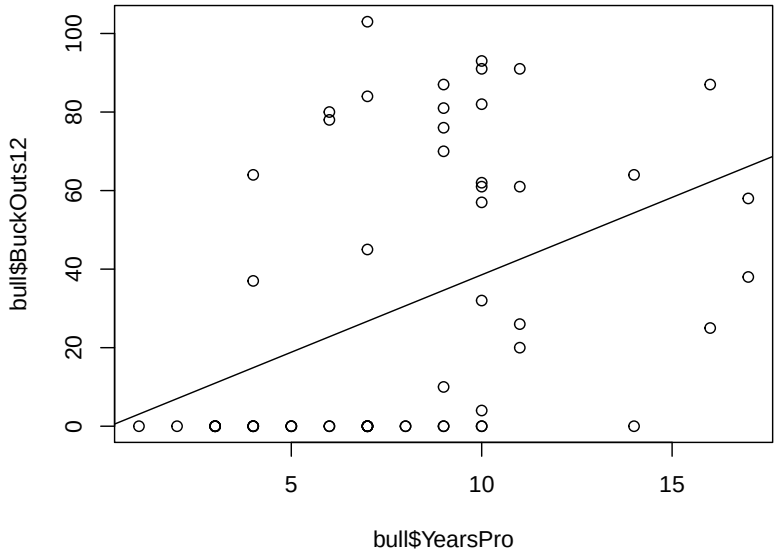



And if you do, we can submit that line. And in our plot window, we see a nice scatter plot of the YearsPro variable against buck outs for the 2012 season. So we can see that the bull riders in this data set have been pro for at least one year, and some as many as more than 15 years. And that throughout the 2012 season, they had buck outs ranging from 0 – up to 100.

Now, what we can also see is that there isn't a real clear linear relationship here. It's just kind of a random scatter or cloud of points, meaning that a change in YearPro doesn't really give us any information about how buck outs will change. So when we see something like this, we're expecting a correlation that's close to 0, or a very weak correlation. And also, if we had to fit a linear model to this data, meaning a line of best fit, it would be pretty much flat.

Now, we can ask R to give us this nice visual of how the line would fit through the data with another function. And that function is called "**abline()**". So **abline()** is going to produce over our scatterplot a model to the

variables. Now because correlation is all about linearity – the linear relationship between two variables – we're going to use the "**lm()**" function, which is a linear model function that we'll learn about when we go over linear regression. And in order to input these variables into the **lm()** function, we're actually going to have to reverse the order. So we want to predict buck outs as a function of YearsPro.




```
abline(lm(bull$BuckOuts12 ~ bull$YearsPro))
```

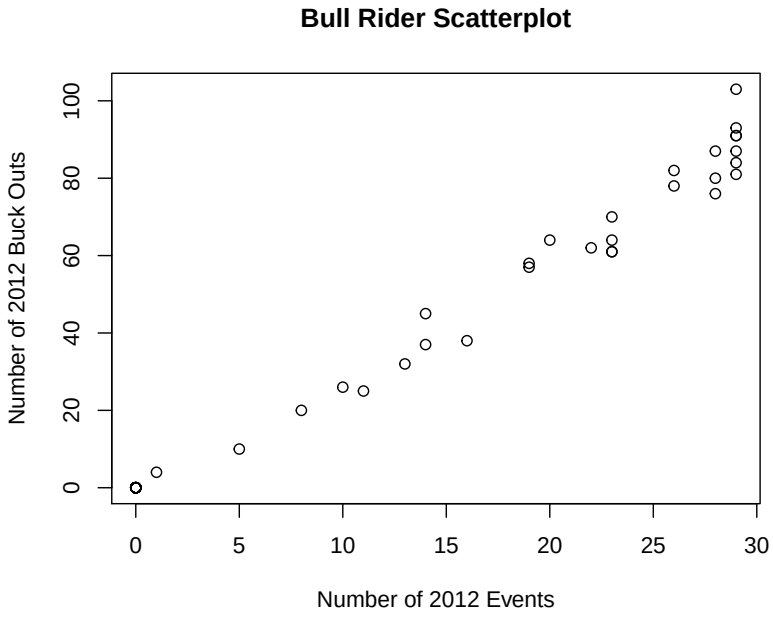
Now, to R, this tilde, or squiggly line, is basically saying as a function of. And so we're giving it the dependent variable, buck outs, which is what we listed second up here, as a function of our independent variable years pro. And this is all again within the "**abline()**" function.

So when we run this, a line fitting through this data. And the points of the data are a little ways away from the line, and the line has a definite slope (it's not flat), indicating a moderate linear relationship.

So to contrast this, let's look at the relationship between another variable and buck outs. Let's look at, instead of YearsPro as our independent variable, let's look at the number of events the riders participated in in 2012.

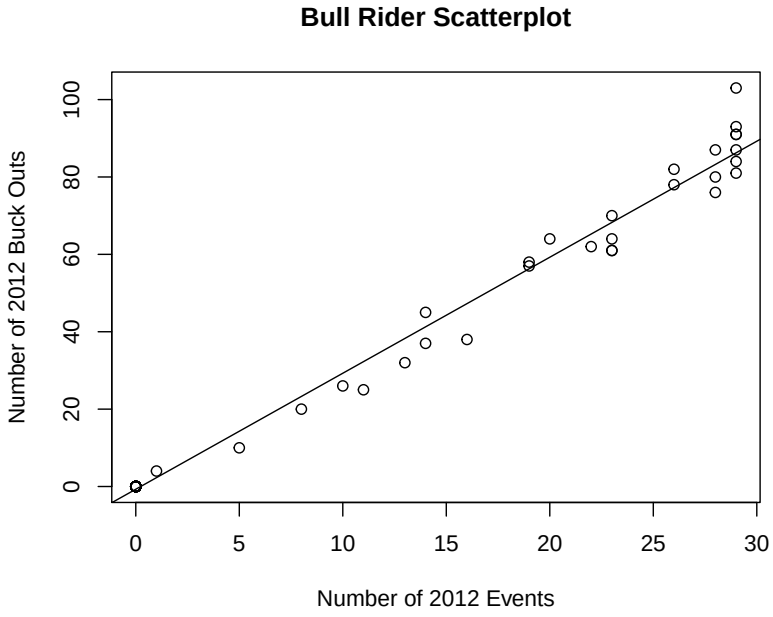
Now if we think about the data set, it would make sense that the more events a rider participated in, the more attempts they would have at riding the bull. So again, I want to plot these, just to confirm.

```
plot(bull$Events12, bull$BuckOuts12, xlab = "Number of 2012 Events",
    ylab = "Number of 2012 Buck Outs", main = "Bull Rider Scatterplot")
```




And I will, of course, add axis labels to any graph I make with these options here.

Again, make sure that closing parentheses is there or else you'll get an error from R. Now if I run this, we're going to get a very different looking scatter plot. So we can see this clear, linear relationship, right? As the writers participate in more events, the number of buck outs that they attempt is definitely going to increase. And that's what we see here confirmed in the scatter plot. It also doesn't really look like there's a non-linear trend. We could fit a line through this data, and it would be pretty much matching what the pattern looks like. So let's just add that line again, with our "**abline()**" function.




```
abline(lm(bull$BuckOuts12 ~ bull$Events12))
```

Again, need to give it the "**lm()**" function within "**abline()**". And then we're just going to give it our variables, but in the reversed order. So we want to say buck outs tilde events.

And here we see that nice line. Most of the points are clustered closely to that line, indicating a strong linear relationship.