Chapter 5: Plotting Data

The ability to communicate your findings to your given audience helps to convince others of the value of your work. Creating simple and effective plots is a vital part of most data analysis projects. In this chapter, we explore various techniques for displaying your data using base R such as scatterplots, histograms, and boxplots. We also mention how to change visual features such as color and titles of your plots.

5.1 Scatterplots

Scatterplots are one of the most popular methods for displaying and summarizing data. A scatterplot is able to display 2 vectors' values on an $x - y$ co-ordinate plane. Each observation's pair of values are simply displayed on this plot using a point. Each point is sequentially plotted to show all of the data succinctly.

To create scatterplot in base R, we can use the `plot()` function. The following example loads the `ToothGrowth` data, obtains some information about the data, then creates a 2D scatterplot.

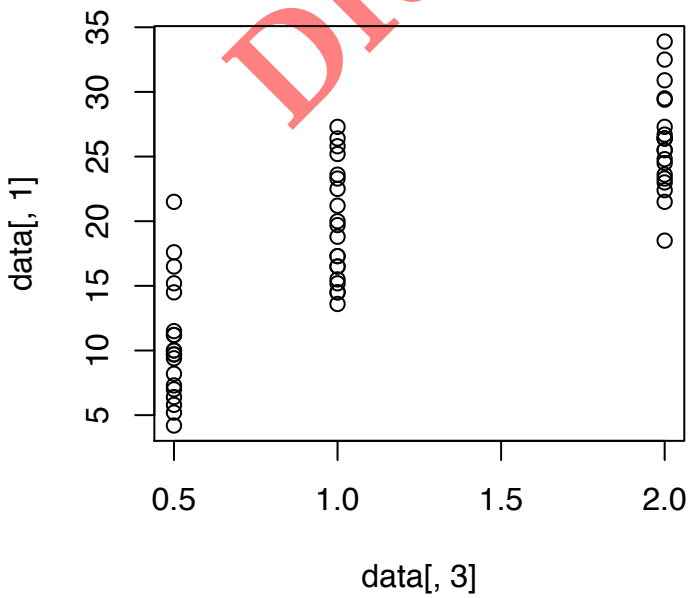
```
#load data
data(ToothGrowth)
#rename to new object
data = ToothGrowth
#get dimension of data
dim(data)

## [1] 60  3
```

```r
#look at top of data
head(data)

##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5

#create scatterplot
plot(x=data[,3], y=data[,1])
```

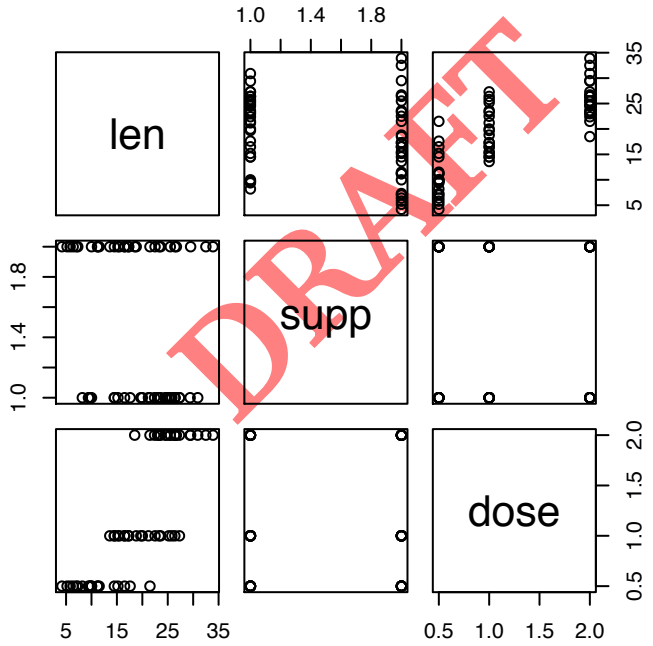
As we can see, the observations tend to have one of three values for the x-axis, while there is decent variation for the y-axis. Further, the range of the x-axis only goes from

0.5 to 2, while the y-axis goes from about 5 to 35. Thus, not only can we obtain a simple scatterplot very easily in R, we are able to gather substantially more insight by inspecting this graphic!

However, we can also create a scatterplot matrix where scatterplots are creating using all pairs of variables. This allows us to quickly gain insight into all of our data very quickly. The following code shows how to create a scatterplot matrix.

```
#create scatterplot

plot(data)
```



To interpret a scatterplot matrix, a singular scatterplot still has only 2 variables. For example, the scatterplot for `len` and `supp` is in the first column and second row box and the second column and first row. Notice that the scatterplots are simply rotated by 90°. Thus, scatterplot matrices are symmetric. Not all of these individual scatterplots have their axis titles directly on the individual scatterplot. For example, you are also able to see the second row's, first column's x-axis values at the bottom of the first column. Identifying

each other individual scatterplot's axis labels follows a similar process for interpretation.

5.2 Adjusting Plots

While producing these scatterplots is very useful, they are not clear to someone who is not familiar with the data. For example, unless you knew what `data[,1]` or `data[,3]` are in the first scatterplot we made, we could not provide a clear interpretation of the data. An obvious solution is to change the titles to clearer descriptions. We might also want to change the color of the points to correspond to different groups. For example, we could want all malignant observations to be red points while benign observations are blue points. The following subsections provide guidance on changing the color, titles, and point shapes of plots.

5.2.1 Color

Color provides an additional dimension of information while still using a 2D scatterplot. Thus, we are able to use 3 variable in on a 2D scatterplot. The option that enables us to do this is `col` of the `plot()` function. Dr. Zheng provides a good reference PDF of a variety of possible colors in R here: `http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf?utm_source=twitterfeed&utm_medium=twitter`.


```
#create scatterplot using color

plot(x=data[,3], y=data[,1], col='cyan')
```