Surveillance Workshop
June-July 2022
Exploratory Plots

Dr Rachel Sippy

June 2022

Instructions

- This is a worksheet intended to help you to create plots in base R to help with exploratory data analysis. If you have any questions or need help, please contact me.
- I will provide hints and some new code in the worksheet to help you make these plots, but I will ask for some additional plots that will require you to write additional code.
- There are many ways to do the same tasks in R, so you can always look at resources online for more ideas to complete the task.
- I recommend you attempt every part of the assignment; even if you do not complete everything.

Plots for Exploratory Analysis

One of the first steps in working with data is to create visualizations to understand what the data look like. This can help us to see if there are any unusual data values (and find potential errors in the data processing steps) and get a general sense for the data. The types of plots that are most helpful or most appropriate will depend on the type of data we have. We have variable types that can be represented by different data formats in R:

***quantitative* variables: these represent some type of number or amount, could be discrete or continuous, may take a positive or negative value or could be a proportion, and may be an integer- or numeric-type variable in R. Examples: body temperature, systolic blood pressure, percent area covered in vegetation**

***qualitative* or *categorical* variables: these represent groups that may be categorical or nominal, they are often a name or could simple be a letter, and may be a factor-, logical- or character-type variables in R. The variable may be binary (e.g. Yes/No, Big/Small), categorical (e.g. districts in a city), or ordinal (e.g. Low/Medium/High)**

Worksheet

Data description

There are two datasets you'll be working on (sent in a previous email):

- `clim.csv`
- `dengue.xls`

Dengue is a viral illness spread by mosquitoes in tropical and sub-tropical countries; mosquito abundance is often driven by climate conditions. The following datasets are from a surveillance study of dengue in a region where dengue is common. Information was collected from clinics, including the total number of daily patient visits during the first four months of the year. We may also have information about the patients who visited. Basic details were collected from patients, including the total number of times they had visited a clinic for acute febrile illness in the past two years.

The datasets contain the following variables:

- `clim.csv`
  - `X` - variable showing the row number for the dataset
  - `Visits` - number of visits to a clinic in a day
  - `Clinic` - clinic ID
  - `Pop` - size of population served by the clinic (in 1000s)
  - `Temp` - mean air temperature (in degrees Fahrenheit)
  - `Prec` - mean humidity (in percentage)
  - `prec` - total daily precipitation (in millimeters)
- `dengue.xls` with sheets `2014` and `2015`
  - `NA.` - variable showing the row number for the dataset
  - `Visits` - number of visits to a clinic by the patient in the past two years

- `Clinic` - clinic ID
- `Age` - age of the patient (in years)
- `Gender` - gender of patient (male, female, nonbinary)
- `SympTemp` - patient temperature at intake (in degrees Celsius)
- `NumSymp` - total number of symptoms reported by patient during the present visit
- `NumChron` - total number of underlying chronic health conditions reported by the patient
- `HistDeng` - patient report of past infection with dengue (1: yes, 0: no)
- `HistDengHouse4Wk` - patient report of recent past infection with dengue (1: yes, 0: no)

Getting started

Create a new folder and make an R Project inside of that folder. Move the dataset files to that folder as well. Open a new R script. Read in both datasets; for the `dengue.xls` file, only read in the sheet called `2014` (*hint* you will need to load a package to help you read one of the files.)

Exploratory plots

Perform the following operations:

`clim.csv`

- Histograms: these are a quick visualization of the distribution of the data. We can assess the range of the data and if it has any skew or extreme values. Histograms are appropriate for *quantitative* data.

```
#Histogram of temperature
hist(clim$temp)

#This will create a histogram with default settings. What are the default #settings?
?hist

#One of the default settings is breaks Try changing this number and see how it
#affects your plot
hist(clim$temp, breaks=10)

#You may notice that the y-axis shows the number of #observations within each histogram bin. This is a
hist(clim$temp, probability=TRUE)

#Adding a normal line: you can calculate a normal #distribution line and add #this to your plot to make
#that will add points to a plot that already exists. Read #about the options for these functions.
x <- seq(min(clim$temp), max(clim$temp), length=length(clim$temp))
f <- dnorm(x, mean=mean(clim$temp), sd=sd(clim$temp))
lines(x, f, col="red", lwd=2)#add a line to show the normal distribution
?lines
?points

#Adding a distribution line: we may want to add a line #that represents the distribution of the data it
hist(clim$temp, probability = TRUE)
lines(density(clim$temp), lwd = 2, col = 'red')
```

- Bar plots: these are a quick visualization of data groups. Bar plots are appropriate for *categorical* data but can also be used with qualitative data.

```r
#Bar plots
#Bar plots in base R require us to calculate the number of #observations in each group before we create
barplot(table(clim$Clinic))

#We may also with to label our plots, which we can easily do:
barplot(table(clim$Clinic), main="Number of Observations by Clinic", xlab="Clinic",ylab="Number of Obse
```

- Scatterplots: this is one way that we may compare how two variables are related to one another. This is appropriate when both variables are *quantitative*.

```r
#Scatter plot of temperature and precipitation
plot(clim$temp,clim$prec)

#We can specify the shape of the dots using the pch #option, and the size using cex. Read the specifics
?plot
plot(clim$temp,clim$prec,pch=12, cex=0.5)
```

- Boxplots: these show the distribution of a variable by group. These are appropriate when one variable is quantitative and one variable is categorical.

```r
#Box plot of precipitation by clinic
#Note that we use a ~ instead of a comma
boxplot(clim$prec~clim$Clinic)

#You may wish to specify that each clinic has a different #color. In this example we simply tell R to u
boxplot(clim$prec~clim$Clinic,col=c(1:10))

#However, you can also use color names or color hexcodes #to specify the colors you want to be used. If
#groups needed, R will simply recycle the colors. See the #attached sheet called "Rcolors.pdf" for refe
boxplot(clim$prec~clim$Clinic,col=c("red","blue","green"))
```

- Line graphs: these are helpful for looking a how a variable changes over time, and can be used *qualitative* data.

```r
#Line graph of of temp
#We will pretend that the temperature data were collected #on a daily basis, and will add a variable to
clim$day<-seq(1,length(clim$temp))#this adds a count variable

#Line graphs are made with plot() and use the type option #to specify 'l' for 'line'
plot(clim$temp,clim$day, type='l')

#We can also show points and lines on the same plot
plot(clim$temp,clim$day, type='b')
```

dengue.xls

- Create a histogram of Age, and add a dotted blue line showing the density of the data distribution. Add appropriate labels for the axes and a title.
- Create a bar plot of Gender, with different colors for each bar. Add appropriate labels for the axes and a title.

- Create a scatterplot to compare of SympTemp and Age, with customized shape, color, and size for the points. Add appropriate labels for the axes and a title.
- Create a boxplot to compare of SympTemp and HistDeng, with customized colors. Add appropriate labels for the axes and a title.