Surveillance Workshop
June-July 2022
Data Preparation

Dr Rachel Sippy

June 2022

Instructions

- This is a worksheet intended to help with your data manipulation skills and preparing datasets. If you have any questions or need help, please contact me.
- Some of the activities, we have already practiced in our R sessions, but some of them will be new. I will provide hints and some new code in the worksheet to help you. The code provided is example code, so you will need to modify it to make it work with your datasets.
- 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.
- **All** of the tasks / analysis should be completed in R. There should be **no** manual manipulation of these datasets (i.e. do not make the changes by opening the file in Excel).
- Please email me your completed worksheet by 20 June 2022.

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 coun-tries; 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.)

Data wrangling

After reading the data in, first step is to clean it for downstream analysis. In particular, perform the following operations:

clim.csv

- Remove the variable showing the row number.

```
#Removing variables
#You can remove a column/variable by indexing (i.e. referring to the number of #the column you want to

#Removing by index
#You can write a command with the numbers of the columns you want to keep.
#Remember, when you use indexing, you need brackets [] after the name of the
#dataset, and the comma inside the brackets allows us to refer to rows or
#columns in the dataset. Anything before the comma refers to row numbers and
#anything after the comma refers to column numbers.
#Here we create a new dataframe that includes all the rows and columns 1-5
#from data1
data2<-data1[,1:5]
data3<-data1[10:20,]#includes rows 10-20 from data1

#You can also exclude columns by using a negative sign
#Here we create a new dataframe that includes all the columns except column 6
data4<-data1[,-6]

#Removing by name
#Here we use the function subset() to remove the column named "Var1"
data4<-subset(data1,select=-Var1)#the negative sign means we are removing
```

- Change the name of the columns to the following names:
  - `temp` - change to `Temperature`
  - `prec` - change to `Precipitation`
  - `Prec` - change to `Humidity`

```
#Renaming variables
#The easiest way to rename a variable is simply to create a new variable
#from the old variable and give it the new name you want. Then remove the old #variable. Remember, to w
#the name of the variable within that dataset.
data1$newname<-data1$oldname
```

- Convert `Temperature` to a degrees Celsius.

```
#Converting variables
#We can convert variables by combining two variables into one variable:
data1$sum12<-data1$num1 + data1$num2

#We can also transform our variables by performing a calculation.
data1$var2<-(data1$var1+50)/(10/3)
```

- Change the clinic ID variable into a factor, with the factor levels in alphabetical order.

```
#Changing variable type
#There are functions to change variable types:
?as.numeric()
```

```
?as.character()
?as.factor()

data1$charnum<-as.character(data1$num1)

#To change to a factor, you must start with a character variable. Remember, we
#use factor for variables that are categorical.
#You must also include information on the levels (i.e. what order to use for
#the different categories)
data1$season2<-as.factor(data1$season,levels=c("Spring", "Summer",
                                               "Autumn", "Winter")))
```

dengue.xls

- Remove the column showing the number of rows in the dataset.
- Change the variables related to the history of dengue into factor variables with levels called `Yes` and `No`. Make sure these correspond correctly according to the description in the dataset.

Examine the data

Answer the following questions about the datasets.

clim.csv

- What is the lowest temperature in Celsius?
- Which clinic has the most observations?

```
#Examining data
#Creating tables can be a useful way to look at how many observations are in
#each category
table(data1$factorvariable)

#You can also cross-tabulate:
table(data1$factorvar1, data1$factorvar2)
```

dengue.xls

- What is the mean age for patients?
- What is the highest temperature reported for a patient?
- How many female patients report a recent history of dengue?
- Which clinic has the lowest mean number of visits?

```
#Examining data
#When you want to summarize data according to categories or groups, you can
#use a function called aggregate(). You need to include information on what #variable to aggregate, wha
#information you want. The aggregate variable can be one or more variables, but must be numeric. The gr
```

```r
#Here we look at the sum of all the variables called num1 within each group
#of another variable called factorvar1
data5<-aggregate(data1$num1,list(data1$factorvar1),sum)

#Here we look at the minimum of each variable called num5, grouped by two #variables
data6<-aggregate(data1$num5,list(data1$factorvar1, data1$year),min)
```