Importing Data in R

Jose Toledo Luna

2024-07-21

Table of contents

Comma-Seperated Values (CSV)

> ⚠️ **Warning**
>
> This tutorial assumes the data set is in working condition. That is we assume the default settings for `read.csv`. In some cases we may need to change the header, specify the field separator and more. See `?read.csv` for further details and examples.

We will now import a `csv` file, to do this we will use the `read.csv` function. A simple template to follow is

```r
read.csv(file = 'path where csv is located in your computer')
```

An easy way to find the location of your data *(or any file)* is using the `file.choose()` function in R. `file.choose()` will bring up a file explorer window that allows you to interactively choose a file path to work with.

In your console, run the following command

```r
file.choose()
```

For example, after running the above command the `births` dataset is located in

```
[1] "/Users/jtoledo/Desktop/Projects/csuf-math-338/data/births.csv"
```

> ⚠️ **Warning**
>
> Depending on your location and operating system the directory will be different

Therefore, to read the `births` data set I would copy/paste the directory of the csv location and run the following command

```r
birth_dat <- read.csv(file = "/Users/jtoledo/Desktop/Projects/ucla-stats13/data/births.csv")
```

```
  X Gender Premie weight Apgar1 Fage Mage Feduc Meduc TotPreg Visits    Marital
1 1   Male     No    116      9   28   34     6     3       2     10    Married
2 2   Male     No    126      8   30   18    12    12       1     14  Unmarried
3 3   Male     No    161      8   28   29    12    12       3     14    Married
4 4   Male     No    133      9   26   23     8     9       3     10    Married
5 5 Female     No    119      8   30   19    12    12       2     12  Unmarried
6 6   Male     No    110      9   30   26    12    16       2     13  Unmarried
  Racemom Racedad   Hispmom   Hispdad Gained      Habit MomPriorCond BirthDef
1   White   White   Mexican   Mexican     30  NonSmoker         None     None
2   White Unknown   NotHisp   Unknown     50     Smoker At Least One     None
3   White   White OtherHisp OtherHisp     65  NonSmoker         None     None
4   White   White   Mexican   Mexican      8  NonSmoker         None     None
5   Black Unknown   NotHisp   Unknown     20  NonSmoker         None     None
6   Black Unknown   NotHisp   Unknown     32  NonSmoker         None     None
     DelivComp BirthComp
1         None      None
2         None      None
3 At Least One      None
4 At Least One      None
5         None      None
6         None      None
```

We are not just limited to `csv` files, we can import data from Excel (in `csv`, `XLSX`, or `txt` format), `SAS`, `Stata`, `SPSS`, or others. A good reference to import various data formats can be found on [datacamp r-data-import tutorial](#)

Text File (TXT)

Next, we consider importing a `.txt` file. To do so we will use the `read.table` function instead of the `read.csv` function. For this example, we consider the `ozone.txt` file from our course website

A simple template to follow is

```
read.table(file = 'path where txt file is located in your computer')
```

After running `file.choose()` on our console and locating the path in which we stored our data

```
file.choose()
```

```
[1] "/Users/jtoledo/Desktop/Projects/csuf-math-338/ucla/stats10/data/ozone.txt"
```

we can copy/paste the path as follows

```
ozone_dat <- read.table(file = "/Users/jtoledo/Desktop/Projects/csuf-math-338/data/ozone.txt"
                        header =TRUE)
```

```
ozone_dat
```

```
          x       y    o3
1 -120.0258 34.4622 0.044
2 -119.7413 36.7055 0.081
3 -121.7333 36.4819 0.035
4 -119.2908 36.3325 0.080
5 -117.1289 32.8364 0.053
```

You will notice we now used an additional argument `header = TRUE` in our `read.table` function. We use `header=TRUE`, whenever the text tile contains names of the variables as its first line.

If we forget to use `header=TRUE`, the first line of the text file will be treated as a row of the dataset and `read.table` will automatically create the variable names for us

```
wrong_ozone_dat <- read.table(file = "/Users/jtoledo/Desktop/Projects/csuf-math-338/data/ozor
```

```
wrong_ozone_dat
```

```
          V1       V2     V3
1          x        y     o3
2 -120.0258 34.4622  0.044
3 -119.7413 36.7055  0.081
4 -121.7333 36.4819  0.035
5 -119.2908 36.3325   0.08
```

In the above example, `read.table` automatically create the variable names `V1,V2,V3` for each column and the first row has values `x,y,o3` *(which is incorrect)*.

In conclusion, some text files do not have variable names in the first row and only contain the actual data. As a result, it is our responsibility to import the data in a suitable manner.