R Crash Course - Loading data into R

Willson Gaul
willson.gaul@ucdconnect.ie

January 2021

Introduction

The first step of data analysis in R is often to load data into R. This document presents four practice tasks to help you practice loading data into R. Each task requires you to download a different dataset onto your computer and then load the data into R. The third and fourth tasks are more difficult and require you to overcome some common challenges that arrise when loading data into R.

Before you start these tasks, you should create a new folder on your computer. You can name the folder anything you want. You will save all the files for these practice tasks in your new folder.

Task 1

Difficulty: *Easy*

1) Download the file 'UCD_tufted_ducks.csv' and save it on your computer in the folder you created for these practice tasks.
2) Create a new R script, name it "*Task_1.R*", and save it in the folder you created for these practice tasks.
3) Set your working directory to the folder you created for these practice tasks.
4) Copy and paste the following line of code into your R script:

```
duck_data <- read.csv("UCD_tufted_ducks.csv")
```

5) Run that line of code to read the data from the file *UCD_tufted_ducks.csv* and save it in R's working memory as an object called `duck_data`. R will look for the file *UCD_tufted_ducks.csv* in whatever folder is currently set as the working directory. If you did not set the working directory to be the same folder that the data file is saved in, R will not be able to find the file and will produce an error message.
6) Type and run the command `head(duck_data)`. When you run this command, the first few rows of the data frame should print in the console. It should look something like this:

```
head(duck_data)
```

```
##    Location       Date   Protocol Tufted_Duck_count
## 1 Main Lake 2018-02-21 Stationary                8
## 2 Main Lake 2018-03-13 Stationary                9
## 3 Main Lake 2018-03-22  Traveling                5
## 4 Main Lake 2018-04-04 Stationary                9
## 5 Main Lake 2018-04-21  Traveling                2
## 6 Main Lake 2018-04-28  Traveling                6
```

Task 2

Difficulty: *Easy*

1) Download the file 'traffic.csv' and save it on your computer in the folder you created for these practice tasks.

2) Create a new R script. Name it anything you want, and save it in the folder you created for these practice tasks.
3) Make sure your working directory is set to the folder you created for these practice tasks. Use `getwd()` to see what the working directory is currently set to, and `setwd()` to change the working directory if needed.
4) Copy and paste the following line of code into your R script:

```
traffic_data <- read.csv("traffic.csv")
```

5) Run that line of code to read the data from the file *traffic.csv* and save it in R's working memory as an object called `traffic_data`.
6) Look at the first few rows of the dataframe using the `head()` command (see Task 1 for an example). The first few rows of `traffic_data` should look like this:

```
##   date_ddmmyyyy time_hhmm vehicles pedestrians notes
## 1    01/05/2020     07:12       19          16    NA
## 2    01/05/2020     09:30       40          13    NA
## 3    01/05/2020     08:46       40          24    NA
## 4    01/05/2020     10:09       35          32    NA
## 5    21/04/2020     16:49       42          58    NA
## 6    21/04/2020     17:36       40          27    NA
```

Task 3

Difficulty: ***Medium***

1) Download the file "plant.txt" and save it on your computer in the folder you created for these practice tasks.
2) Open the file "plant.txt" in a text editor program (for example, Notepad or notepad++ on Windows, TextEdit or BBedit on Mac OSX, gedit on Linux, or others). If you do not know how to find a text editor, ask now!
3) Take a look at the "plant.txt" file in your text editor, keeping an eye out for things that might be important when you read the data in to R. Then, close the text editor.
4) Read the data from the file "plant.txt" in to R.
5) How many rows and columns are there in the data that you read into R? Use `str()` or `dim()`. You should find **40 rows** and **2 columns**. If you do not have 40 rows and 2 columns, try changing how you read in the data so you get the correct number of rows and columns. Aks for help if you need it.
6) What are the names of the columns? Modify the following line of code so that it uses the name of your data frame, and then paste the code into your R script and run it:

```
colnames(plant_df)
```

Your output should look like this:

```
## [1] "treatment" "height_cm"
```

7) In your R script, write a short comment about how to read the data from "plant.txt" in to R (use `#` to start comment lines).

Task 4

Difficulty: ***Hard***

1) Go to the data repository at http://doi.org/10.5281/zenodo.3964574.
2) Download the data file "003Corrected_PointCounts_PtAbbaye2019.csv" and save it to your computer. This file contains data from bird point counts conducted in Michigan, USA, in 2019.
3) Read the data from "003Corrected_PointCounts_PtAbbaye2019.csv" into R.
4) Look at the first few rows of the data using `head()`.
5) How many rows and how many columns are in the data frame? (Hint: use either `str()` or `dim()`. You should find 848 rows and 21 columns).
6) The column named "`count`" gives the number of birds of each species that were counted during each survey. What data type has R used to store the data in that column? (Hint: use `str()`). Is R using the data type that you expect? If not, why not?