USE R! HAVE FUN!

Yuan Tian

2016年4月18日

Hey guys, I hope you all enjoy last week's exercises. I think that's enough for you as your point of departure in R programming. And during this week, we will learn something about getting and cleaning data. In my point of view, this is the vital step in data science because one need to manipulate the data into a useable form. So do pay attention about this week's exercise.

Let's get down to business.

Chapter 2 Obtainning and cleaning data

The pipeline of solving a problem with data science

1. Define the question
2. Choose the ideal dataset
3. Determine what data can be accessed
4. Obtain data
5. Clean the data
6. Do some exploratory data analysis
7. Apply mathematical tools(e.g. statistical prediction)
8. Interpret results
9. Challenge results
10. Synthesize results
11. Create reproducible code
12. Distribute the results

The steps above are the basic pipeline when dealing with a problem using data science. And we will cover step 4 to 11 during our tutorials. But what we learn with this tutorial is far from enough as being a data scientist. So I hope you can take this course as a point of departure.

Getting data

Set your working directory

Before you get down to real analysis, I highly suggest you to check your working directory and set your working directory properly.
You can check you working directory using the following code.

```
getwd()
```

And you can redirect your working directory using the following code.

```
setwd("your ideal path")
```

Read local flat files

Before reading your data into R, make sure it is in your working directory. While you can search what is in your working directory by hand, you can also check that out using the following code.

```
dir()
```

You are all set up when you make sure you have all the data you need in your working directory. And the following functions are two handy tools reading local flat files.

```
read.table("file name", header = TRUE, sep = ";")
read.csv("file name")
```

Have a glance at your data

After loading your data, you should take a look at them first. And you can use the following code to do so.

```
head(dataset)
str(dataset)
dim(dataset)
```

Try this with the following code.

```
library(datasets)
head(airquality)
str(airquality)
dim(airquality)
```

You will see a dollar sign "$" in the result. And this operator is a link between a dataset and a certain column. I mean if you want to manipulate certain column of the data, you should specify which column it is and to which dataset it belongs.

Subset

This might be a function that you may want to use.

```
subset(the dataset, your condition)
```

Exercises

1. You did your first R program in last chapter, but don’t worry, most of functions that you can ever imagine have been finished by others, so you just need to find them and use them. Install your first R package by typing “install.packages(”ggplot2“)”

2. There is a handy function in ggplot called “qplot”. Type “?qplot” to see the help documentation.

3. Load the default dataset called “airquality” using the following code.

```
library(datasets)
```

How many columns are in this dataset.

4. The airquality dataset have 5 months’ records from 5 to 9. What is the average of solar rediation(Solar.R) each month?

5. Make a plot to illustrate the average solar rediation each month.