USE R! HAVE FUN!

Yuan Tian

2016-4-25

Hey, guys. We have learned about the basic operations of R and basic steps of data science. But don't rush to formal analysis. Because it is really important to have a glance at data first to determine if the data is suitable for analysis. And that's exactly what exploratory data analysis does.

Game time.

Chapter 3 Make graphs with R

Why graphs

Exploratory tools can help us sharpen our hypothesis to boost our efficiency in data analysis. And the whole point of making plots is to help us understand the properties of data. I quote some tips from the course "Exploratory Data Analysis" (https://www.coursera.org/learn/exploratory-data-analysis/home/welcome) offered by JHU.

1. Show comparisioins of your data.
2. Show causality, machanism, or systemic structure.
3. Show multivariate data.
4. Integrate all evidences you have.
5. Describe and document the evidence with appropriate labels, scales, and so on.
6. The content is the king.

Plotting systems in R

There are 3 plotting systems in R, namely the base plotting systems, tne lattice system, and the ggplot2 system. The characteristics of each systems is listed in the following:

The base plotting system

- The process is intuitative, you start with a blank plate, then add each features of the plot.
- All the features can be customized, so it's convient.
- Once you made a mistake, you cannot go back. I mean it.

The lattice plotting system

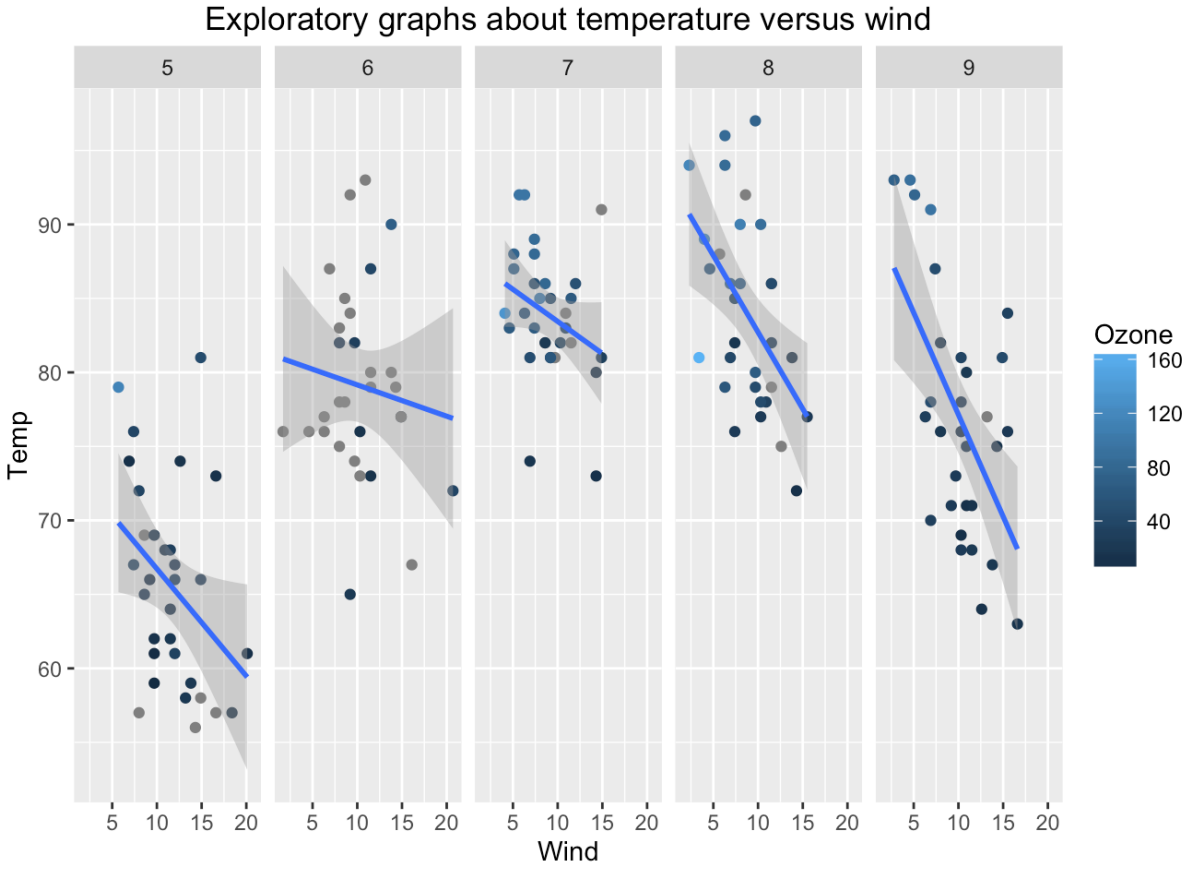
- All the plots are made with one single command, so the process is not that intuitative.
- But it is extremely suitable for conditioning types of data.
- Many things are adjusted by the program, which means you may find it hard to customize a plot.

The ggplot2 plotting system (gg: grammar of graphs)

- This is a mixture of the previous 2 systems.
- Things are adjusted by the program, but you can also customize it in easily.
- In my point of view, this system offers us with graphs in a way prettier manner.

GGPLOT2 system

The “gg” stands for grammar of graphs, which means we will have certain definitions of every element that will appear in a plot. By specifying all the elements, we will make a plot at our will. We will only cover this plotting system in our tutorials because your energy is limited. I want to make your tiem count.
I will demonstrate how to make a plot like the following one:



Basis elements in ggplot2

- ggplot: stands for the data set, which is your “paletee”.
- aes: short for aesthetic mapping, you can assign some properties of your data such as color, size, etc.
- geom: short for geometric objects like points, lines, and shapes.
- facets: used to make multi-panel plot.
- stats: short for statistical transformation.
- scales: used to specify what scale map should use. (eg., red for male and blue for female)
- labs: used to specify the labels and titles.

Step 1 Make a paletee

```
library(datasets)
p <- ggplot(airquality, aes(Wind, Temp))
print(p)
```

You will find nothing showed in the screen, that's because we just made a palette but did not put anything on that. So that's blank.
Here, by using ggplot() function, we made a palette where all the data will come from the "airquality" dataset. And we use aes() to specify that is Wind and y is Temp.

Step 2 Add points

```
p <- p + geom_point(aes(color = Month, alpha = 0.5))
print(p)
```

Here I used a function belonging to geom, the geom_point function, to add points on the palette we made. There are also other functions to specify geometric features such as geom_smooth. I also used the aes function to sepcify more characteristics of the plot. You can search their meanings by yourself.

Step 3 Split the plot into different panels

```
p <- p + geom_point(aes(color = Month, alpha = 0.5)) + facet_grid(.~Month)
print(p)
```

Here I used a funciton belonging to facet to make a multi-panel plot. By ".~", I set other commands with default options.

Step 4 Add labels

```
p <- p + geom_point(aes(color = Month, alpha = 0.5)) + facet_grid(.~Month) +
 labs(title = "Exploratory graphs about temperature versus wind") + geom_smo
oth(method = "lm")
print(p)
```

We can use labs() function to add legends to the plot. The parameter "title" allows you to add a title, "x" allows you to add the labels of x-axis, and similar for "y".
Good luck with your exercises.

Exercises

1. With the airquality dataset, is there any relationships between the solar readiation (Solar.R) and Ozone? Support your analysis with proper graphs.
2. Make a box plot to compare temperatures from different months.
3. Find the mode of wind, and support your analysis with proper graphs.