Intro to Social Science Data Analysis

Lecture 6: Data Visualisation in R

Christopher Gandrud

October 2, 2012

1. Assignment 2
2. Recap
3. Principles of Graphics Excelence

Assignment 2

Due: Friday 19 October

Describe at least **3** variables in a data set.

You need to select a **range of descriptive statistical tools**. The tools should include both **numerical descriptive statistics** and **graphics**.

These tools should describe the variables':

- central tendency,
- variation,
- their relationships with the other variables.

The descriptions need to be discussed **in paragraph form**.

The description must be **reproducible**. So you should email me the link to a Dropbox folder with:

- the `.csv` data set,
- the `.Rmd` R markdown file,
- the final `.html` file.

Quick Quiz (1)

When you describe data, what **two** things do you always need to discuss?

Why do you need to describe both things?

Give examples for data at different measurement levels.

Quick Quiz (2)

What is the difference between the **population** mean and the **sampling** mean?

Why would you log transform a variable?

Today

Last week: we largely learned how to describe our data *numerically*.

Today: we will learn how to present our data with *graphics*.

We will learn both how to create graphics in R, but also the principles of effective statistical graphics.

Inferential Statistics

Many of the things we learn today will also apply to inferential statistics.

The first part of this lecture is based on Tufte (2001)

Many of the examples are from the Junk Charts Blog (`http://junkcharts.typepad.com/`**).**

We will also use the World Bank data we downloaded last class.
R Source Code at: `http://bit.ly/OTWEGS`

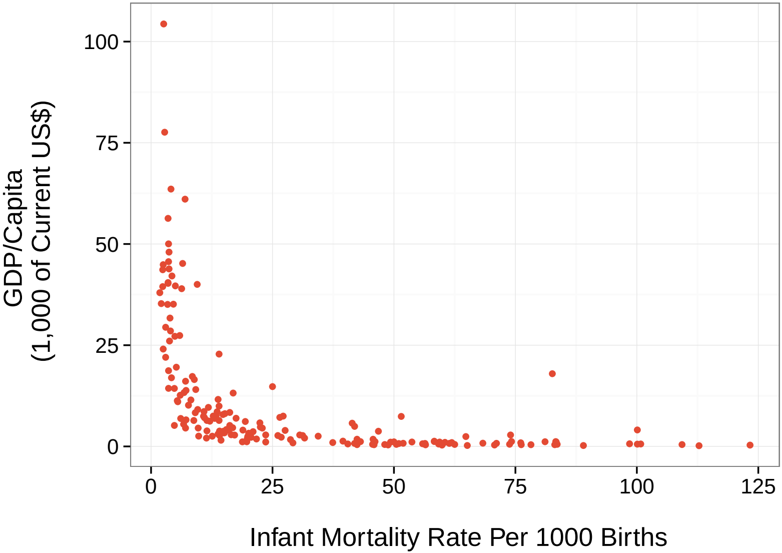
Why Graphics?

Why use graphics? Why not just describe all of our data in tables?

```
# Create data frame with GDP/Capita & Infant Mort.
DataDump <- InfantNoMiss[,
              c("GDPperCapita", "InfantMortality")]

# Show data
DataDump

##     GDPperCapita InfantMortality
## 7        38959.8             6.3
## 8          425.1            76.2
## 9        13829.8             7.2
## 10        3795.7            14.1
## 11        2803.3            17.2
## 12        4068.5           100.1
## 13        7665.1            13.4
## 15       45638.1             3.6
## 16       42101.4             4.3
## 18        4950.3            41.9
## 19        4534.1             7.1
## 20       13181.3            16.9
```


GDP/Capita
(1,000 of Current US$)
100
75
50
25
0
0
25
50
75
100
125
Infant Mortality Rate Per 1000 Births

Avoid Circles! (1)

In general: Avoid using the *size* of a circle to mean something!

Avoid Circles! (2)

Why?

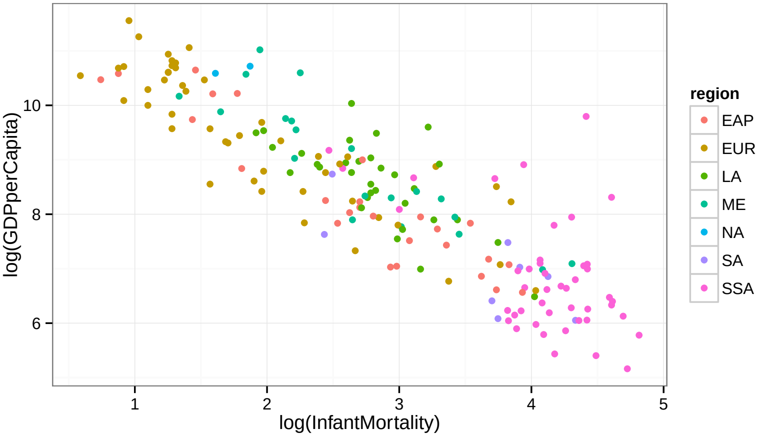
Colours! (1)

Colours

There are a number of ways to specify colours in ggplot2.

The simplest way is to let ggplot choose the colours for you.

```r
# Create scatter plot divided by region
ggplot(data = InfantNoMiss, aes(log(InfantMortality),
                                log(GDPperCapita),
                                colour = region)) +
      geom_point() + theme_bw()
```



Professional Graphics With R

Many people use R to create professional graphics.

For example: see the New York Times' graphics blog: `http://chartsnthings.tumblr.com/`

They often use R in combination with Adobe Illustrator.

See Nathan Yau's Book *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics* (`http://book.flowingdata.com/`).

References I

Tufte, Edward R. 2001. The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press.