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
4. Base R Graphics
5. ggplot2

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/`).

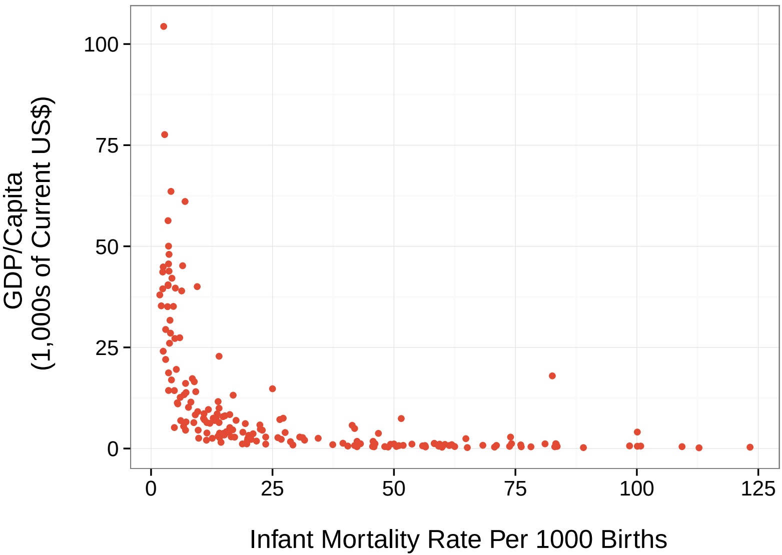
We will also use the World Bank data we downloaded last class.
R Source Code at: `http://bit.ly/OTWEGS`

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

Goal of Statistical Graphics:

The efficient communication of complex quantitative ideas.

Tufte's Principles for Excellent Statistical Graphics (2001, 13) (a selection):

- show the data
- encourage the eye to compare differences in the data
- serve a clear purpose
- avoid distorting the data
- be closely integrated with the text

Tufte's Principles

Tufte's Principles for Excellent Statistical Graphics (2001, 13) (a selection):

- show the data
- encourage the eye to compare differences in the data
- serve a clear purpose
- avoid distorting the data
- be closely integrated with the text

Tufte's Principles for Excellent Statistical Graphics (2001, 13) (a selection):

- show the data
- encourage the eye to compare differences in the data
- serve a clear purpose
- avoid distorting the data
- be closely integrated with the text

Tufte's Principles

Tufte's Principles for Excellent Statistical Graphics (2001, 13) (a selection):

- show the data
- encourage the eye to compare differences in the data
- serve a clear purpose
- avoid distorting the data
- be closely integrated with the text

Tufte's Principles for Excellent Statistical Graphics (2001, 13) (a selection):

- show the data
- encourage the eye to compare differences in the data
- serve a clear purpose
- avoid distorting the data
- be closely integrated with the text

Show the Data

Encourage the eye to compare differences.

Serve a purpose.

Show the data, not other things like silly graphics, or unnecessary words.

Have high **data ink** ratio.

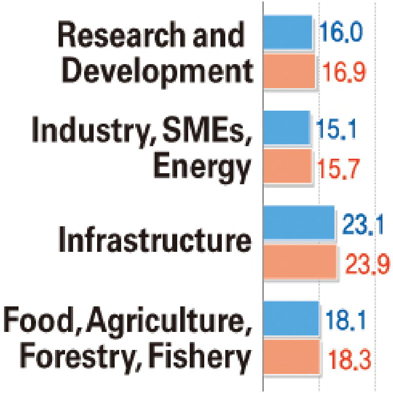
$$\text{Data Ink Ratio} = \frac{\text{data} - \text{ink}}{\text{total ink}} \tag{1}$$

Example

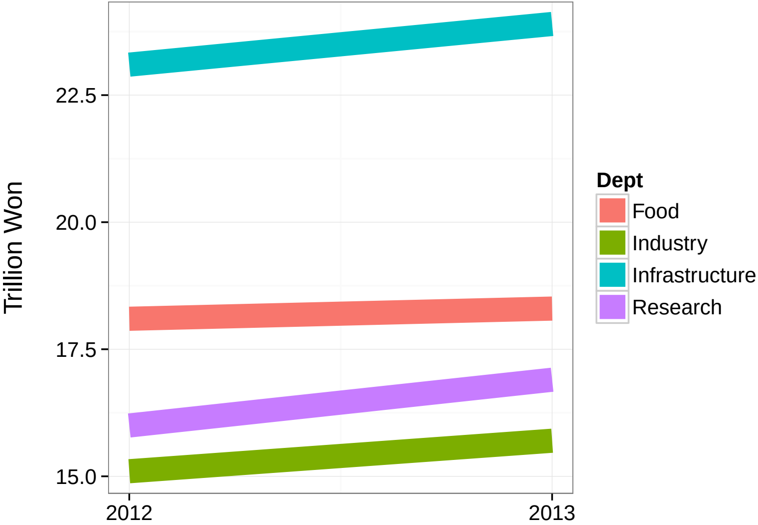
Encourge the Eye to Compare Differences

How did the budgets change?

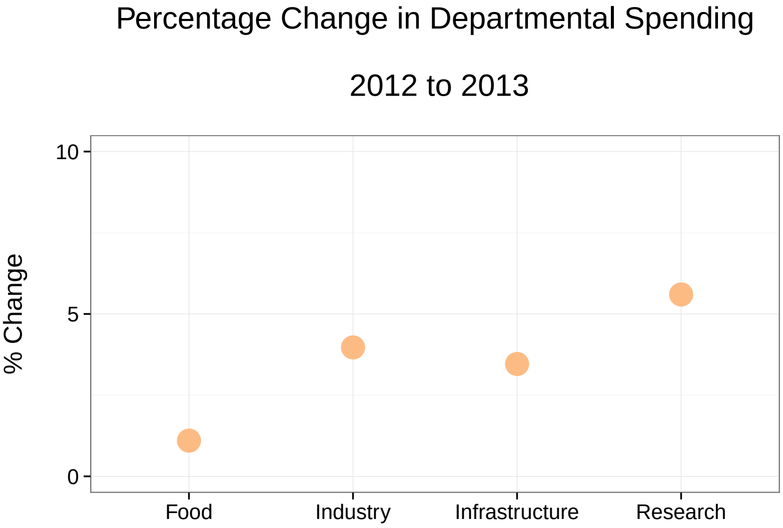
(Orange is 2013, Blue is 2012)



A Little Better



Even Better


Percentage Change in Departmental Spending
2012 to 2013
% Change
0
5
10
Food
Industry
Infrastructure
Research

Avoid distorting the data.

Special case: Circles.

Avoid Circles! (1)

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

This includes avoiding:

- Bubble charts
- Pie charts

Avoid Circles! (1)

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

This includes avoiding:

- Bubble charts
- Pie charts

Avoid Circles! (2)

Why?

Circles can distort data.

- It is very difficult to compare their size
- The Ebbinghause Illusion!

Avoid Circles! (2)

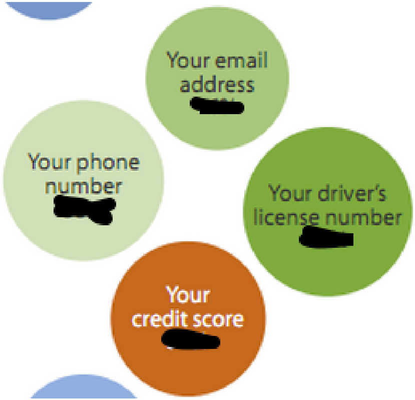
Why?

Circles can distort data.

- It is very difficult to compare their size
- The Ebbinghause Illusion!

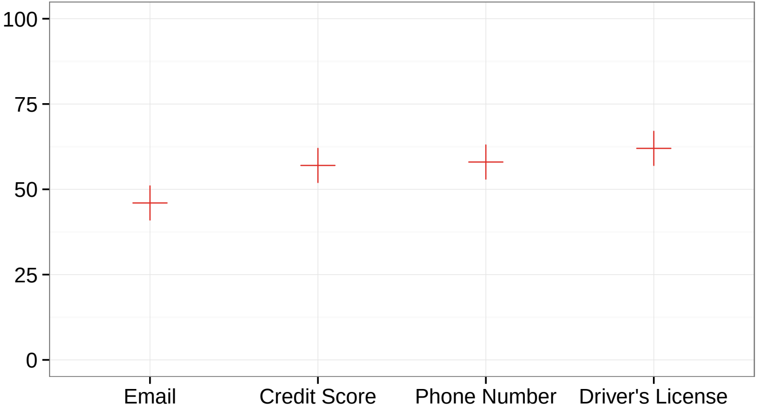
Order the 4 circles from largest to smallest.

The circles are on a scale of 0-100, so how much bigger are each of the circles relative to each other?

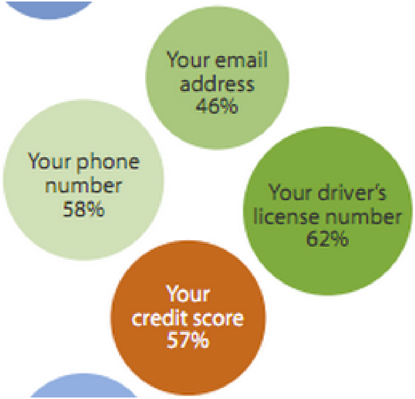


Order the 4 bars from largest to smallest.

How much bigger are each of the circles relative to each other?

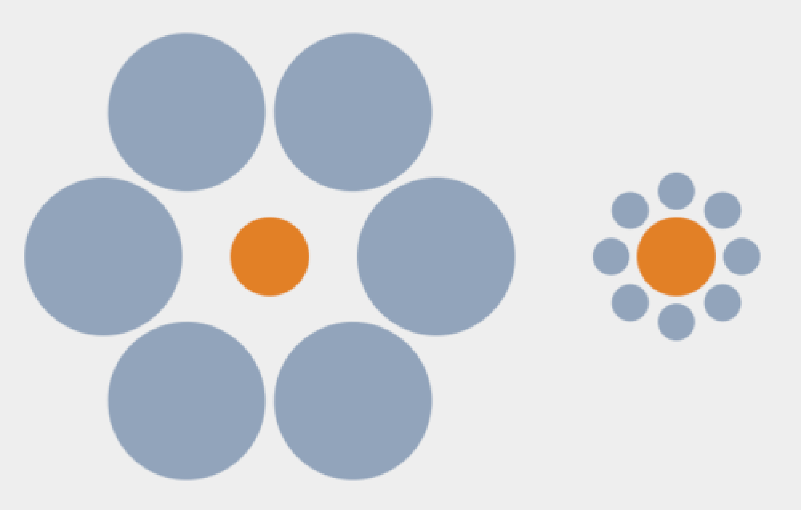


The circles basically tell you nothing that a simple table could do better.



Ebbinghause Illusion

Which circle is bigger?



Another Issue

Colour blindness.

Colour blindness:

People who are colour blind can have difficulty distinquishing between red-green and blue-yellow.

About 5-8% of men are colour blind.

We need to choose colour schemes for our graphics that are **colour blind friendly**.

For more information see `http://www.usability.gov/articles/newsletter/pubs/022010new.html`.

Warning

Remember:

Graphics are only as good as what you put in them.

Silly data and statistics will always create silly graphs.

Base R Graphics

Last week we saw that R has some basic graphics functions like:

- `plot`
- `histogram`

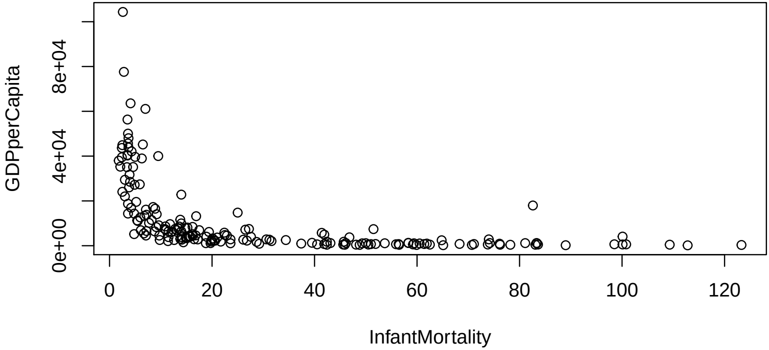
Base R Graphics

Last week we saw that R has some basic graphics functions like:

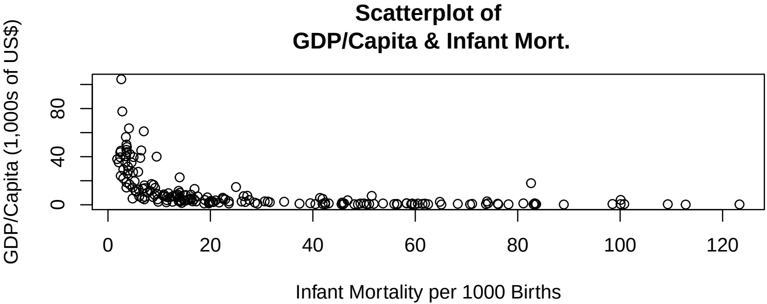
- `plot`
- `histogram`

A Basic Scatter Plot

```
# Create a basic scatter plot
with(MortalityGDP,
     plot(x = InfantMortality, y = GDPperCapita))
```



```r
# Basic scatter: axis labels & rescale GDP/Capita
with(MortalityGDP,
     plot(x = InfantMortality,
          y = (GDPperCapita/1000),
          xlab = "Infant Mortality per 1000 Births",
          ylab = "GDP/Capita (1,000s of US$)",
          main =
          "Scatterplot of\n GDP/Capita & Infant Mort."))
```



Graphics with ggplot2

The ggplot2 package

The ggplot2 package allows us to do much more than base R graphics.

What is ggplot2?

“gg” means “Grammar of Graphics”

(2 just means it is the second one)

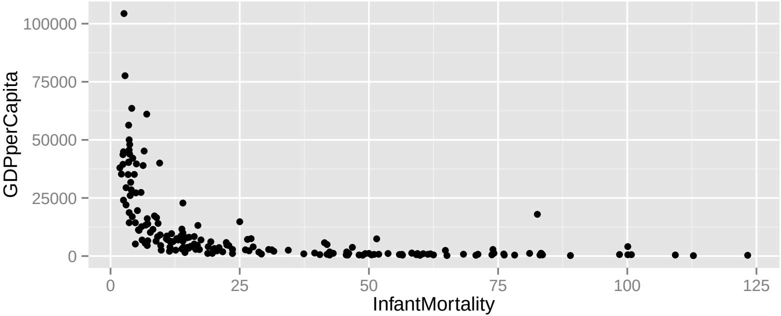
Each plot is made of **layers**. Layers include the coordinate system (x-y), points, lables, etc.

Each layer has **aesthetics** (`aes`) including the x & y, size, shape, and colour.

The main layer types are called **geometrics** (`geom`). These include lines, points, density plots, bars, and text.

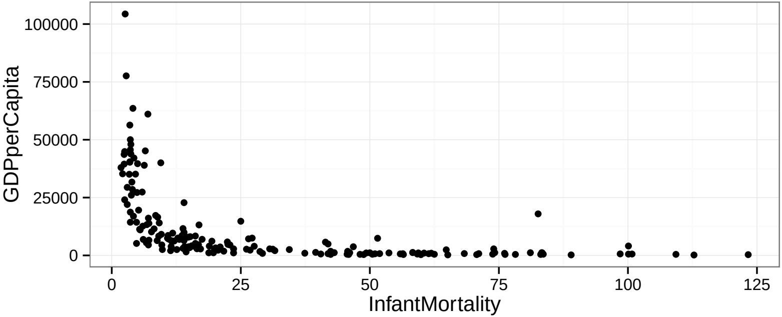
Simple Example

```
# Scatterplot
ggplot(data = MortalityGDP, aes(x = InfantMortality,
                               y = GDPperCapita)) +
    geom_point()
```



Simple Example with Blank Theme

```
# Scatterplot
ggplot(data = MortalityGDP, aes(x = InfantMortality,
                                y = GDPperCapita)) +
     geom_point() +
     theme_bw(base_size = 13)
```



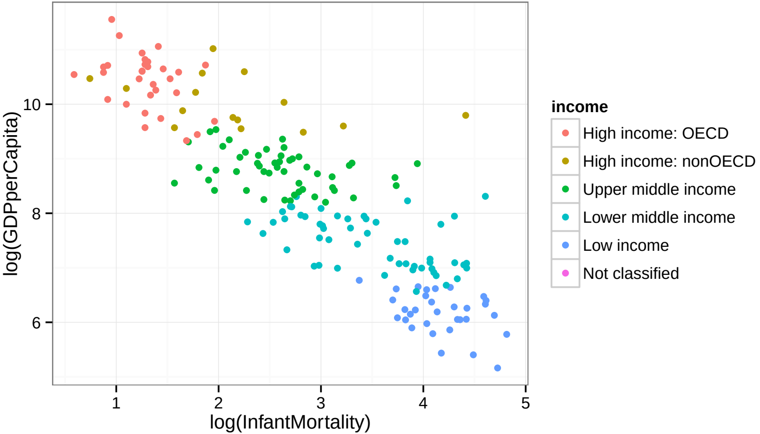
Colours! (1)

Colours

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

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

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



Colors! (2)

There are many ways to pick specific colors.

In this class we will mainly use **hexadecimal** colours.

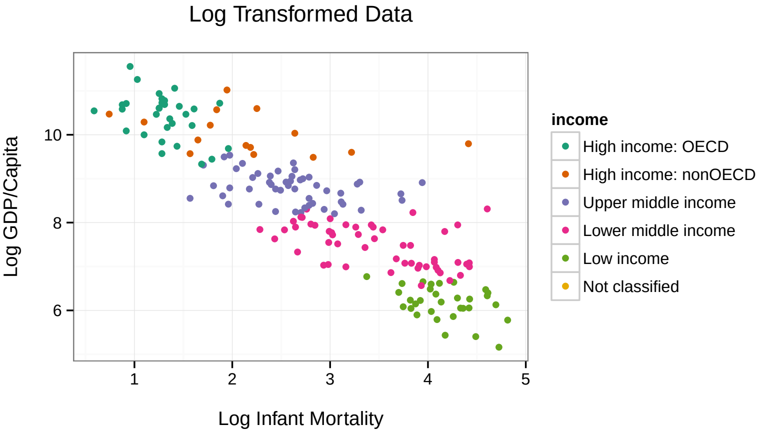
This is probably the most comonly used system for choosing colours on the web.

Every colour is given six digits.

A good website for getting hexadecimal colour schemes is: `http://colorbrewer2.org/`.

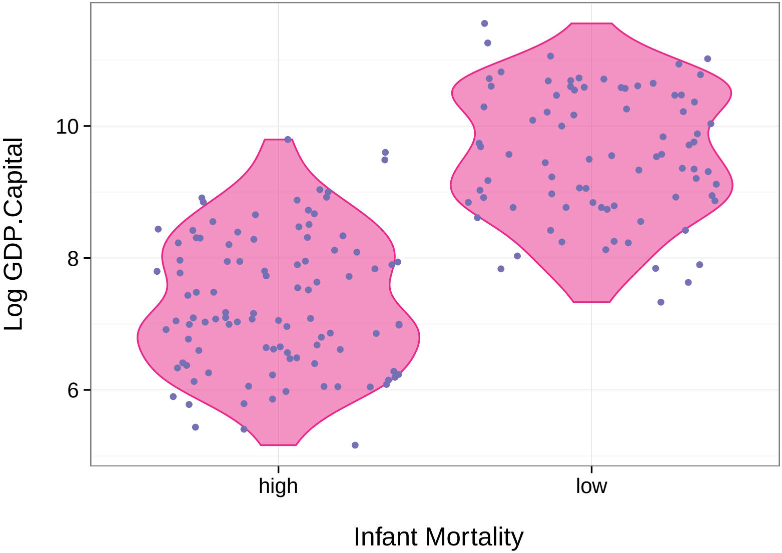
```
# Create colour vector
Colours <- c("#1B9E77", "#D95F02", "#7570B3",
             "#E7298A", "#66A61E", "#E6AB02")

# Create graph
ColourScatter <- ggplot(data = InfantNoMiss,
                        aes(log(InfantMortality),
                            log(GDPperCapita))) +
                  geom_point(aes(colour = income)) +
                  scale_color_manual(values = Colours) +
                  xlab("\nLog Infant Mortality") +
                  ylab("Log GDP/Capita\n") +
                  ggtitle("Log Transformed Data\n") +
                  theme_bw()
```

```
# Show scatter plot
ColourScatter
```



```r
# Create a violin Plot
Violin <- ggplot(InfantNoMiss, aes(
                                factor(DumMort),
                                log(GDPperCapita))) +
                  geom_violin(fill = "#E7298A",
                              colour = "#E7298A",
                              alpha = I(0.5)) +
                  geom_jitter(color = "#7570B3") +
                  xlab("\n Infant Mortality") +
                  ylab("Log GDP.Capital\n") +
                  theme_bw(base_size = 16)
```

```
# Create a violin Plot
Violin
```



More Info

More Information:

http://docs.ggplot2.org/current/index.html

Maps with GoogleVis

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.