Visualizing Relationships

Statistics 4868/6610 Data Visualization

Prof. Eric A. Suess

2/22/2016

Introduction

Today we will go over what correlation measures and some of the examples from Chapter 6.

Examples from the book

The plots from Chapter 6

- Scatterplot
- Scatterplot Matrix
- Correlogram
- Bubble Chart

Correlation is not Causation

In introductory Statististics courses the difference between **Correlation** and **Causation** is discussed. These two ideas are not the same.

Often it is said, "Correlation is not causation."

Correlation

The **Correlation Coeficient**, r, measures the *strength* and *direction* of the *linear association* between two *quantitative variables*.

Memorize this!

Good interview question.

Causation

One variable **causes** an effect, linear or non-linear, on another another variable.

Causality

Probabilistic Causation

Confounding variables

A confounding variable is another variable that influences the other variables.

Simpson's Paradox

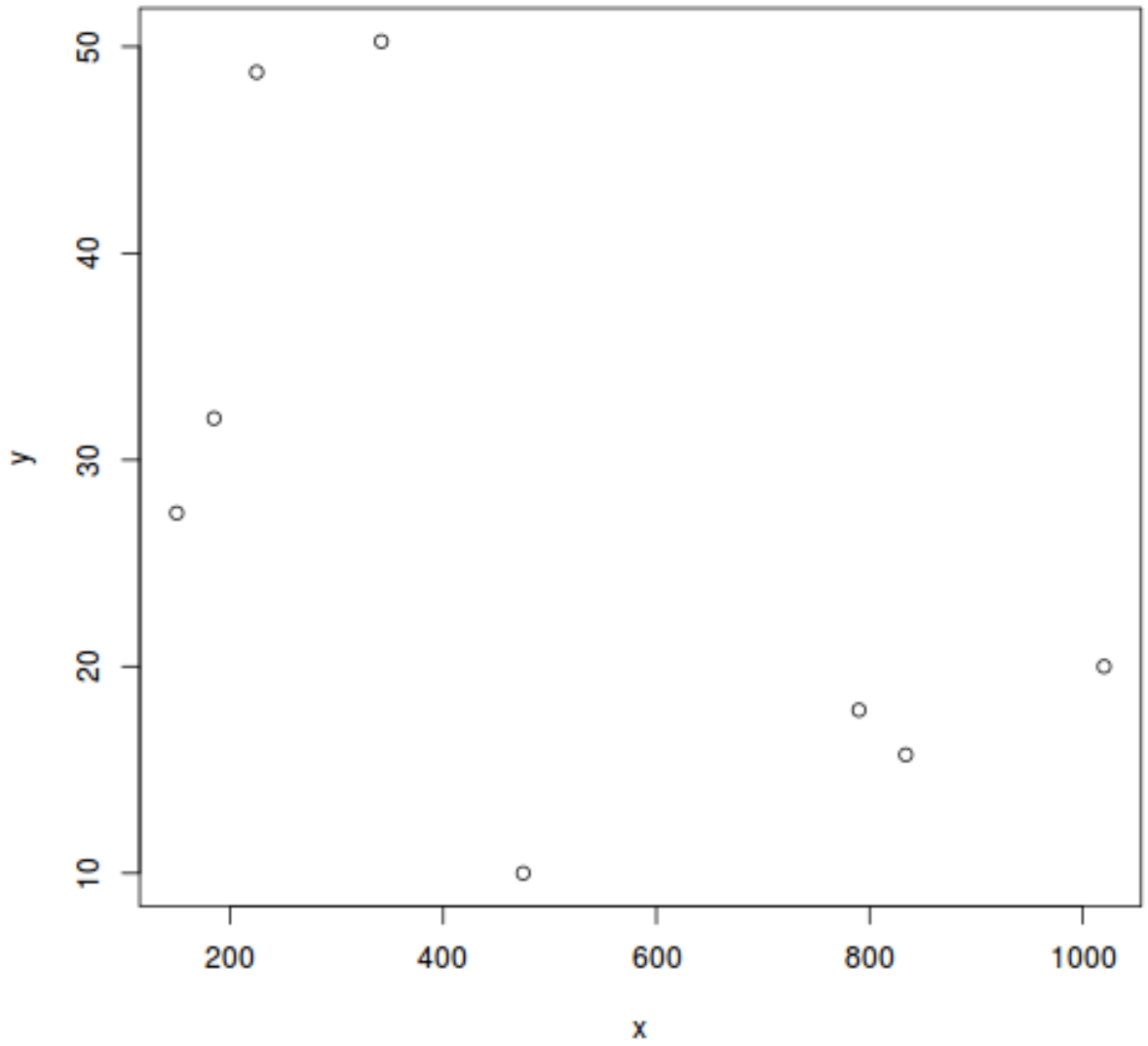
Edward Simpson: Bayes at Bletchley Park

Example of Simpson's Paradox

```
### synthetic data

# Consider book price (y) by number of pages (x)

z = c("hardcover","hardcover",
      "hardcover","hardcover",
      "paperback", "paperback","paperback",
      "paperback")

x1 = c( 150, 225, 342, 185)
y1 = c( 27.43, 48.76, 50.25, 32.01 )

x2 = c( 475, 834, 1020, 790)
y2 = c( 10.00, 15.73, 20.00, 17.89 )

x = c(x1, x2)
y = c(y1, y2)
```

Example of Simpson's Paradox

```
plot(x,y)
```



Example of Simpson's Paradox

```
# correlation

cor(y, x)
```

```
[1] -0.5949366
```

```
cor(y1, x1)
```

```
[1] 0.8481439
```

```r
cor(y2, x2)
```

```
[1] 0.9559518
```

Example of Simpson's Paradox

```r
# linear regression

lm(y ~ x)
```

```

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   41.15238     -0.02665
```

Example of Simpson's Paradox

```r
# linear regression

lm(y1 ~ x1)
```

```

Call:
lm(formula = y1 ~ x1)

Coefficients:
(Intercept)           x1
    13.0613       0.1177
```

Example of Simpson's Paradox

```r
# linear regression

lm(y2 ~ x2)
```

```

Call:
lm(formula = y2 ~ x2)

Coefficients:
(Intercept)           x2
    1.72389      0.01819
```

Example of Simpson's Paradox

Summary: Simpson's Paradox is the changing of the direction of a relationship with the introduction of another variable.

The relationship between Price and Number of pages in a book changes with the introduction of the variable Type of Book (Hardcover, Paperback).

See the R Markdown document SimpsonsParadox available on RPubs.com/esuess.

My favorite plot

The **Scatterplot matix** is a very useful plot for seeing the correlations between variables in a dataset.

Not so useful with more than about 10 variables.

What to do with more variables?

The Correlogram is very useful

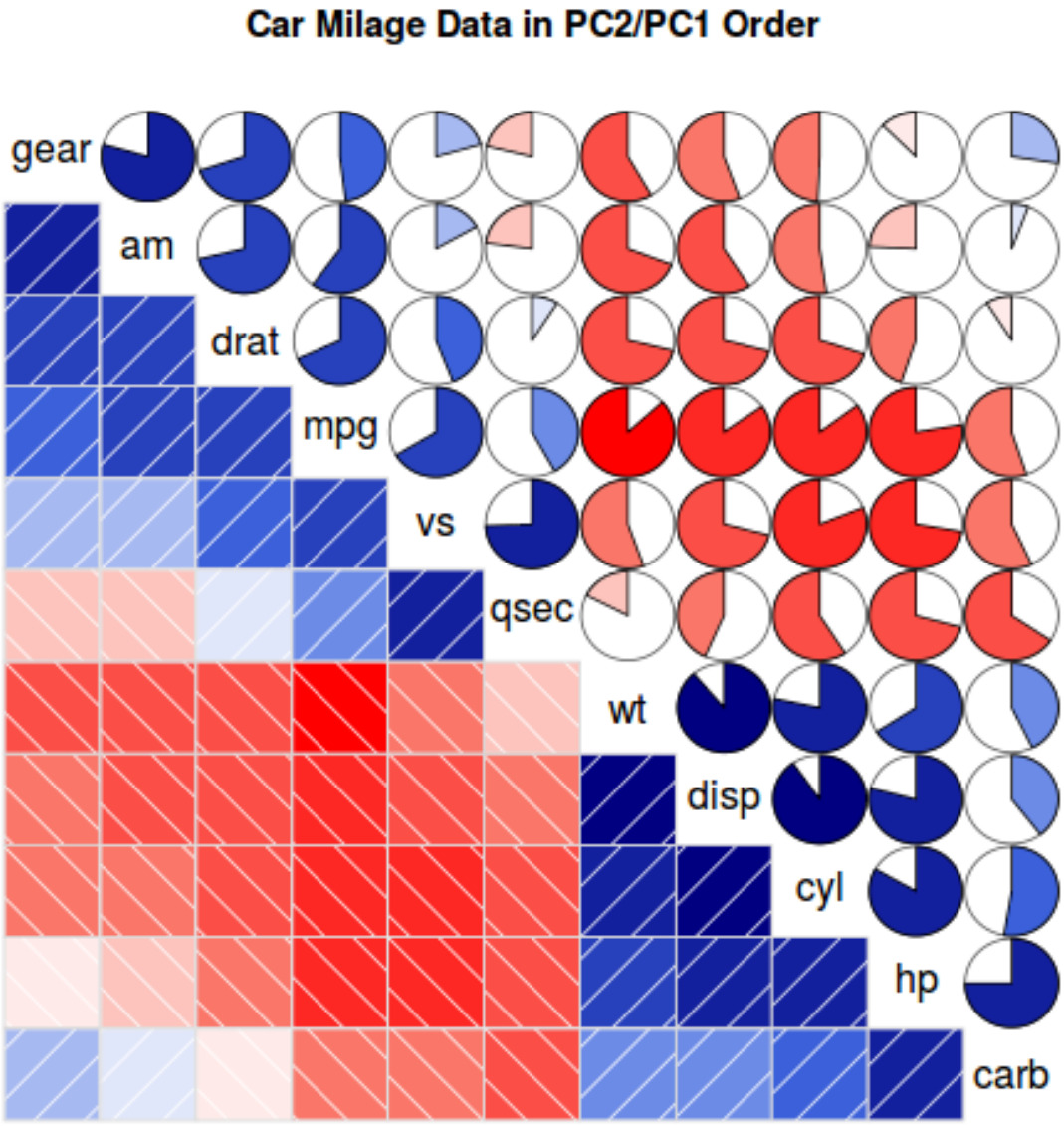
From the Quick-R website.

Correlogram

```
library(corrgram)
```

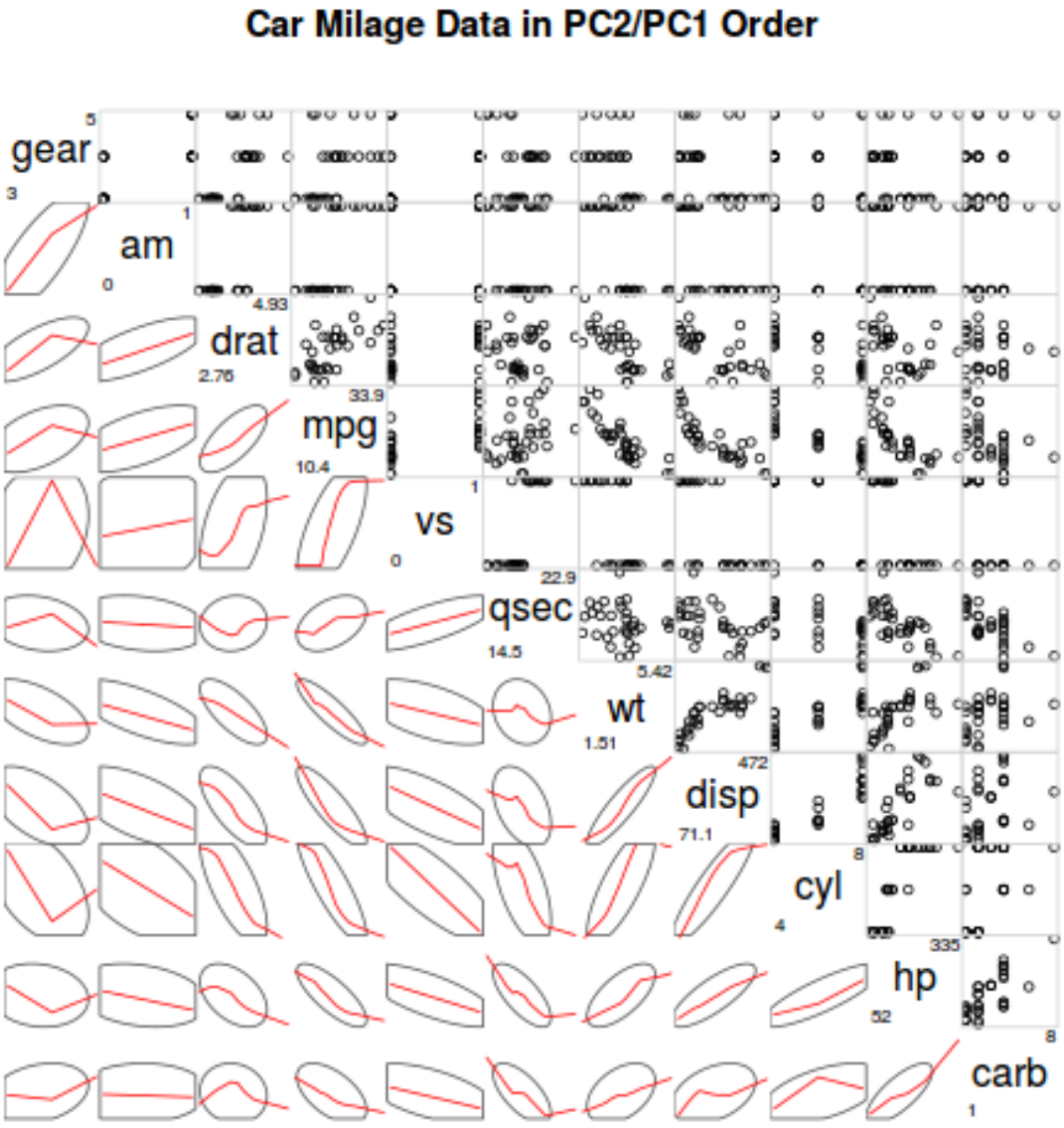
R code

```
corrgram(mtcars, order=TRUE,
  lower.panel=panel.shade,
  upper.panel=panel.pie,
  text.panel=panel.txt,
  main="Car Milage Data in PC2/PC1 Order")
```



R code ==========

```
corrgram(mtcars, order=TRUE,
  lower.panel=panel.ellipse,
  upper.panel=panel.pts,
  text.panel=panel.txt,
  diag.panel=panel.minmax,
  main="Car Milage Data in PC2/PC1 Order")
```



Bubbles

Be sure to study the discussion of the use **area** in the section about **Bubbles** on pages 193 and 194. Look at Figure 6-12, 6-13, 6-14, 6-16 and 6-17. Study Figure 6-17 that uses the correct sized cirles.