Exploratory Data Analysis

Damian Thomas

2017-02-03

Topics

What is Exploratory Data Analysis?

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.[1]

[1] Source: https://en.Wikipedia.org/wiki/Exploratory_data_analysis

History

The history of R is closely tied to EDA

- John Tukey works for Bell Labs
- Bell Labs develops the S language
- John Tukey writes a book: Exploratory Data Analysis, Tukey, (1977).
- S-Plus is commerically successful
- R created – an open source implementation of the S language

EDA is:

- Data focused
- Informal. No model is specified
- Gain insight into the data generating process.
- Learn about the data, underlying structure
- Summarize the data without losing information.
- Gather key information required to build a model.
- Generate questions
- help decide what sort of model fits

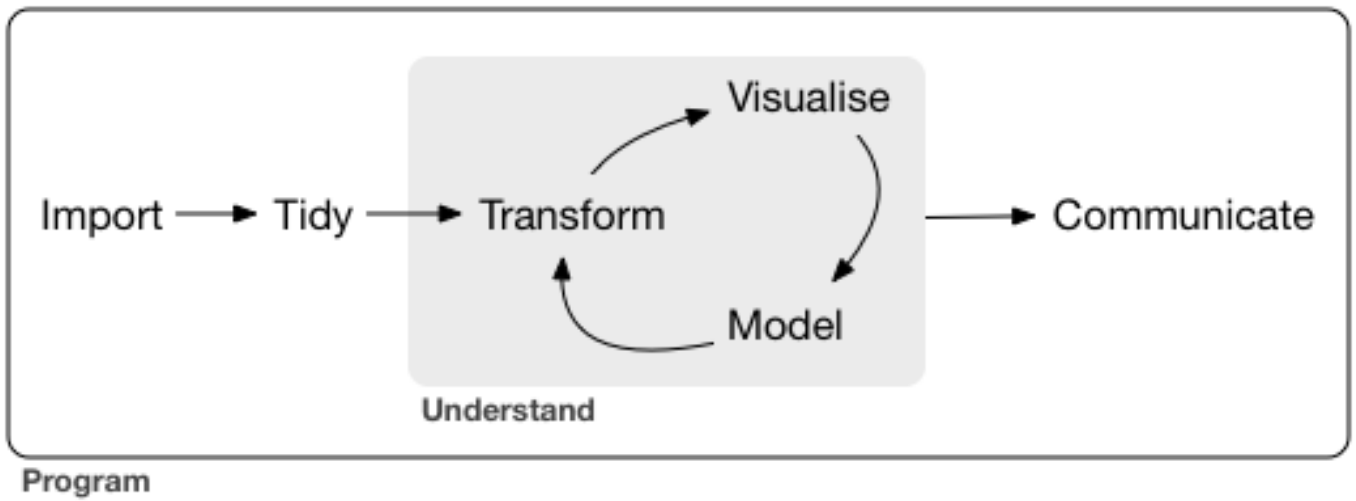
EDA is *not:*

- Model focused
- Dependent on assumptions (randomness, normality, etc.)
- A rigorous formal approach
- Model Specification (regressions, ANOVA)
- Parameter estimation
- Hypothesis testing /statistical inference

Techniques

- Tukey's five number summary (for any distribution)
  - minimum: smallest value
  - lower quartile: 25th percentile
  - median: middle value
  - upper quartile: 75th percentile
  - maximum: largest value
- Summary statistics: characterize the distribution
  - Extremes: range (minimum and maximum)
  - Location: median, mean
  - Spread: quartiles, standard deviation
  - Shape: modality, skew
- Visualizations: Present the data, facilitate discovery
  - Box plot
  - Scatter plot
  - Line plot
  - Bar plot
  - Histogram

The Data Analysis Process[2]



- Write code to carry out each step
- Save the code so you can reproduce (and share) your work

[2]Source: R for Data Science `http://r4ds.had.co.nz/`

Get raw data

Start by getting the data

Tidy Data

Apply tidy principles and restructure your data if necessary

- Each variable should have its own column
- Each observation should have its own row
- Each value should have its own cell

Transform

Create new variables, change existing ones, remove extraneous information.

Model

Use what you've learned to specify a model and apply classical statistics to confirm exporatory findings

Iterate

Continue to refine the analysis. Use what you learn to ask more questions and learn more about the data.

Communicate

Share your results

Example Data Sets: Anscombe's Quartet

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.[3]

[3] Source: `http://en.wikipedia.org/wiki/Anscombe's_quartet`

read.table() family of functions

```
> ?read.table()
> ?read.csv()
> ?read.delim()
```

These functions read raw data saved in delimitied text files, and return a data frame object.

Read raw data from a csv file

```
> anscombe <- read.csv("data/anscombe.csv",
+                      stringsAsFactors = FALSE)
```

Read raw data from a csv file

```
> anscombe <- read.csv("data/anscombe.csv",
+                      stringsAsFactors = FALSE)
> anscombe
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Tidy Data Checklist

- Each value should have its own cell.

Tidy Data Checklist

- PASS: Each value should have its own cell.
  Every cell has a single value.

Tidy Data Checklist

- PASS: Each value should have its own cell. Every cell has a single value.
- Each variable should have its own column.

Tidy Data Checklist

- PASS: Each value should have its own cell.
  Every cell has a single value.

- PASS: Each variable should have its own column.
  Every column is uniquely named and has one kind of data.

Tidy Data Checklist

- PASS: Each value should have its own cell.
  Every cell has a single value.

- PASS: Each variable should have its own column.
  Every column is uniquely named and has one kind of data.
- Each observation should have its own row.

Tidy Data Checklist

- PASS: Each value should have its own cell.
  Every cell has a single value.
- PASS: Each variable should have its own column.
  Every column is uniquely named and has one kind of data.
- FAIL: Each observation should have its own row.
  Anscombe's quartet is comprised four separate data sets. The raw data file gives us all four side-by-side. Therefore each row represents 4 observations.

Tidy Data Steps

```
> rawdata <- anscombe
> for ( q in 1:4 ) {
+     xvar <- rawdata[[ paste("x", q, sep = "") ]]
+     yvar <- rawdata[[ paste("y", q, sep = "") ]]
+     slice <- data.frame(q = q,
+                         x = xvar,
+                         y = yvar)
+     if ( q == 1 ) {
+         anscombe <- slice
+     } else {
+         anscombe <- rbind(anscombe, slice)
+     }
+ }
```

Tidy Data: Result

```
> anscombe
   q  x     y
1  1 10  8.04
2  1  8  6.95
3  1 13  7.58
4  1  9  8.81
5  1 11  8.33
6  1 14  9.96
7  1  6  7.24
8  1  4  4.26
9  1 12 10.84
10 1  7  4.82
11 1  5  5.68
12 2 10  9.14
13 2  8  8.14
14 2 13  8.74
```

Summary Statistics

```
> summary( anscombe[anscombe$q == 1, c("x", "y")] )
```

```
       x                y         
 Min.   : 4.0   Min.   : 4.260  
 1st Qu.: 6.5   1st Qu.: 6.315  
 Median : 9.0   Median : 7.580  
 Mean   : 9.0   Mean   : 7.501  
 3rd Qu.:11.5   3rd Qu.: 8.570  
 Max.   :14.0   Max.   :10.840  
```

Summary Statistics

```
> summary( anscombe[anscombe$q == 2, c("x", "y")] )
       x                y         
 Min.   : 4.0   Min.   :3.100  
 1st Qu.: 6.5   1st Qu.:6.695  
 Median : 9.0   Median :8.140  
 Mean   : 9.0   Mean   :7.501  
 3rd Qu.:11.5   3rd Qu.:8.950  
 Max.   :14.0   Max.   :9.260  
```

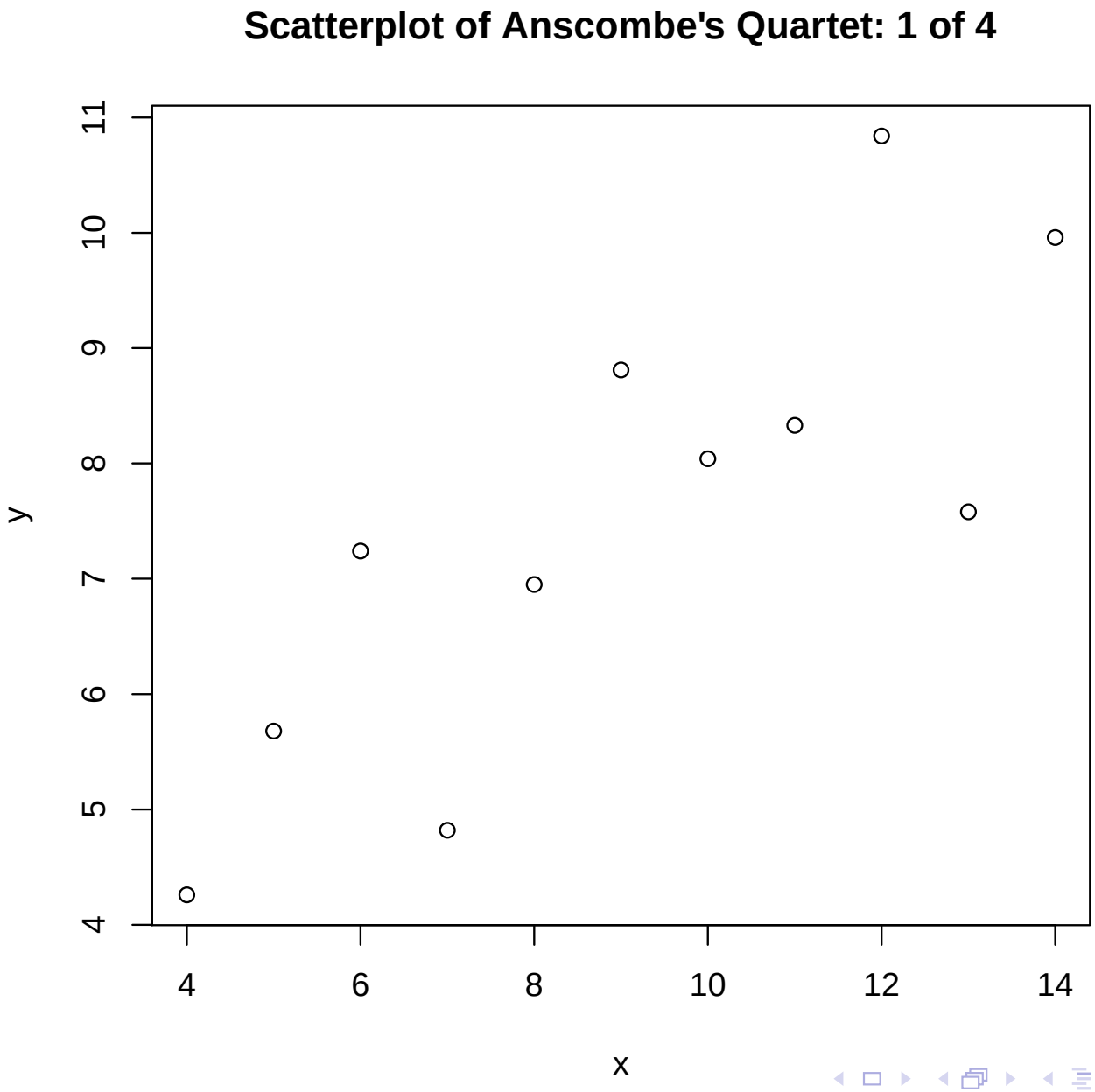
Summary Statistics

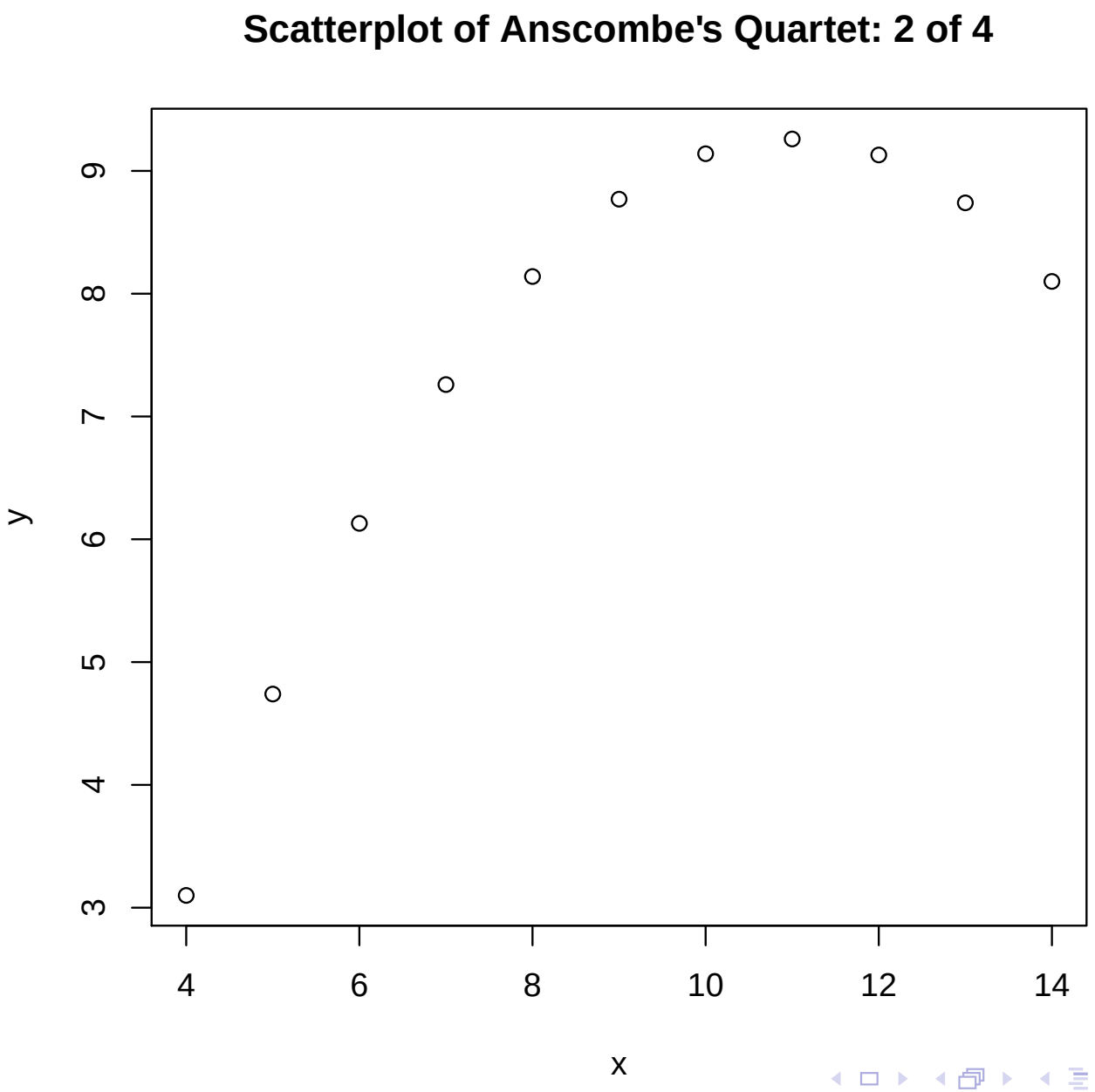
```
> summary( anscombe[anscombe$q == 3, c("x", "y")] )
```

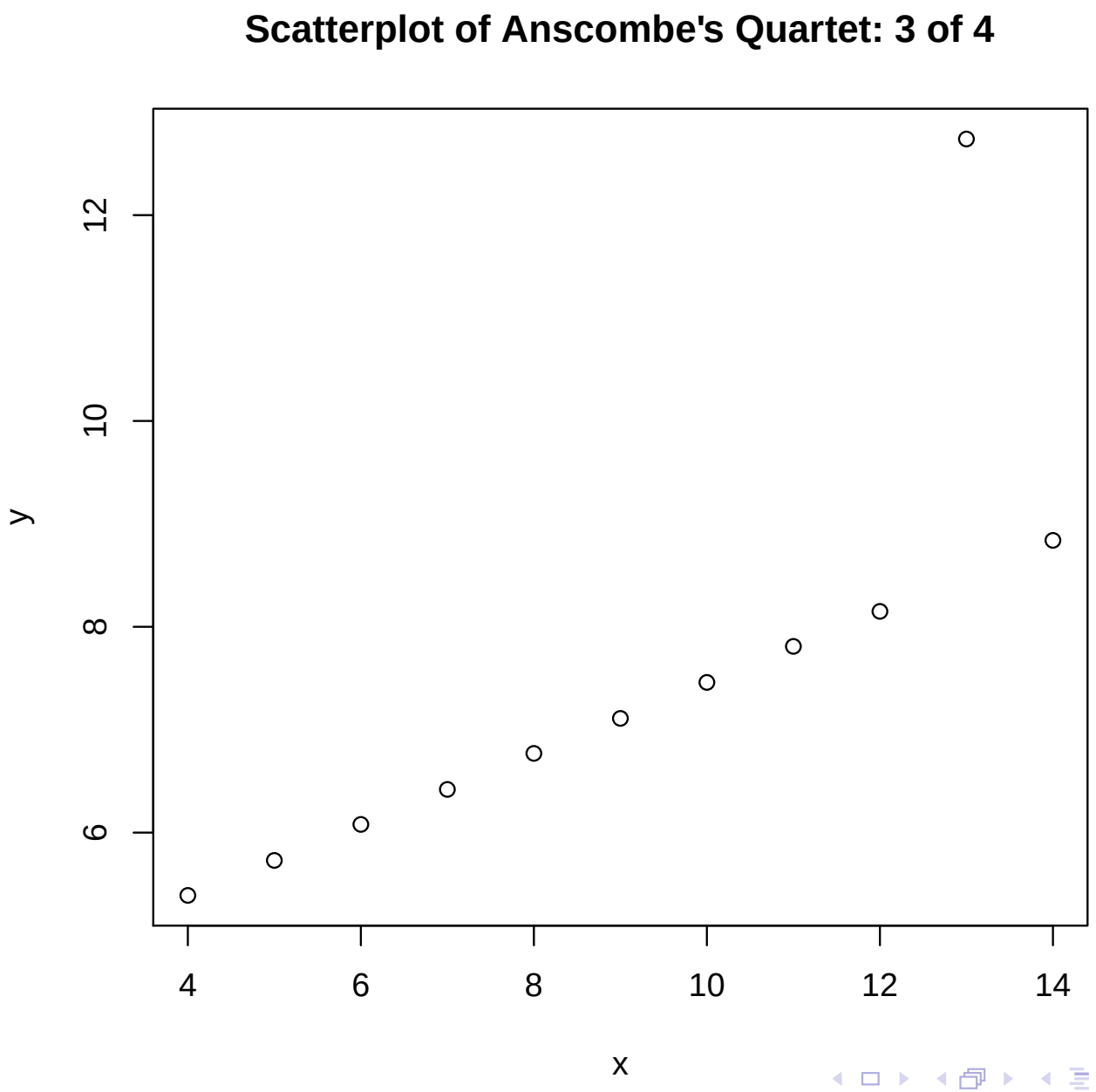
```
       x                y         
 Min.   : 4.0   Min.   : 5.39  
 1st Qu.: 6.5   1st Qu.: 6.25  
 Median : 9.0   Median : 7.11  
 Mean   : 9.0   Mean   : 7.50  
 3rd Qu.:11.5   3rd Qu.: 7.98  
 Max.   :14.0   Max.   :12.74  
```

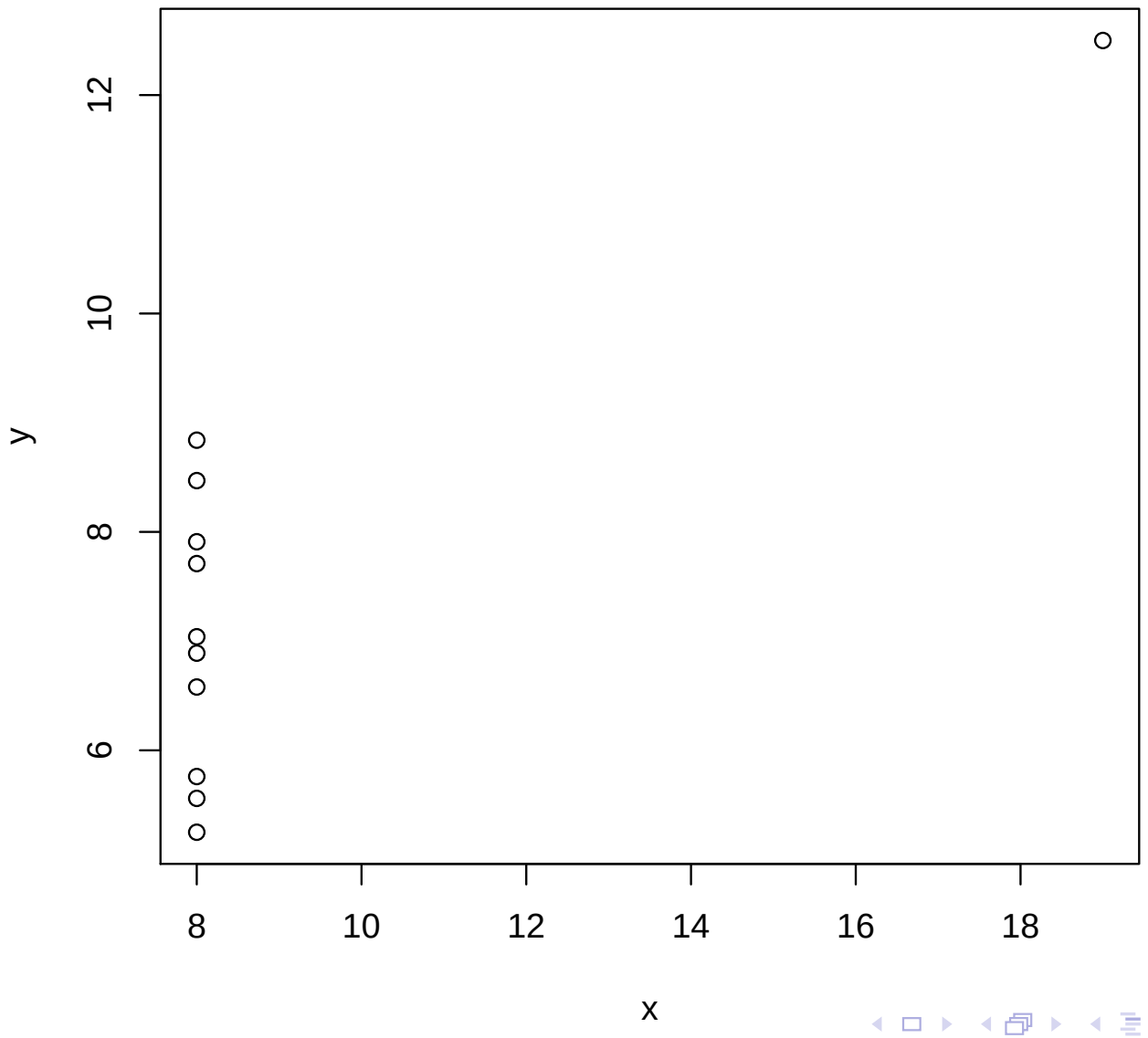
Summary Statistics

```
> summary( anscombe[anscombe$q == 4, c("x", "y")] )
       x                y        
 Min.   : 8   Min.   : 5.250  
 1st Qu.: 8   1st Qu.: 6.170  
 Median : 8   Median : 7.040  
 Mean   : 9   Mean   : 7.501  
 3rd Qu.: 8   3rd Qu.: 8.190  
 Max.   :19   Max.   :12.500  
```


Scatterplot of Anscombe's Quartet: 1 of 4
y
x
4
5
6
7
8
9
10
11
4
6
8
10
12
14


Scatterplot of Anscombe's Quartet: 2 of 4
y
x
3
4
5
6
7
8
9
4
6
8
10
12
14


Scatterplot of Anscombe's Quartet: 3 of 4
y
x
4
6
8
10
12
14
6
8
10
12


Scatterplot of Anscombe's Quartet: 4 of 4
y
x
6
8
10
12
8
10
12
14
16
18