Descriptive statistics and plots

Overview

Plots and statistics of categorical data

Plots and statistics of quantitative data

Announcement: Homework 2

Homework 2 is available

It is due on Gradescope by 11pm on Monday July 14th

- Question 4 involves reading a short article and commented on it, so you can get started on this right away

How did homework 1 go?

Questions/comments about anything?

Review: Vectors

Vectors are ordered sequences of numbers or letters

The c() function is used to create vectors

```
> v  <-  c(5, 232, 5, 543)
> s  <-  c("statistics", "data", "science", "fun")
```

One can access elements of a vector using square brackets []

```
> s[2]        # what will the answer be?
```

We can also apply functions to vectors

```
> sum(v)
```

Review: Data frames

Data frames contain structured data

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review: OK Cupid data

Review: Data frames

Data frames contain structured data

```
> library(SDS111)
> download_data("profiles_revised.csv")      #  only needs to be run once
> profiles <- read.csv("profiles_revised.csv")
> View(profiles)       # the View() function only works in R Studio!
```

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review: Data frames

Variables

Cases

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review: Data frames

Quantitative Variable

Categorical Variable

Cases (observational units)

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review: Data frames

We can extract the columns of a data frame as vector objects using the $ symbol

```
> the_ages <- profiles$age
```

Can you get the sum of the ages of users in this data set?

```
> sum(the_ages)
```

Questions?

Categorical data

Categorical variables

A categorical variable assigns each observation to one of *k* groups

Which variables in the profiles data frame are categorical?

- Is age a categorical variable?

For categorical variables, we usually want to view:

- **Frequency table**: How many items are each category
- **Relative frequency table:** The proportion (or percentage) of items in each category

$$\text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$

Frequency table

```
# Get information about drinking behavior
> drinking_vec <- profiles$drinks

# Create a frequency table showing how often people drink
> drinks_table <- table(drinking_vec)
> drinks_table
```

Relative frequency table

We can create a relative frequency table using the function:

```
> prop.table(my_table)
```

Can you create a relative frequency table for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
> prop.table(drinks_table)
```

Bar plots (pun intended?)

We can plot the number of items in each category using a bar plot

```
> barplot(my_table)
```

Can you create a bar plot for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
> barplot(drinks_table)
```

Let’s try it in RStudio!

Details matter!

Can you figure out how to label the axes?

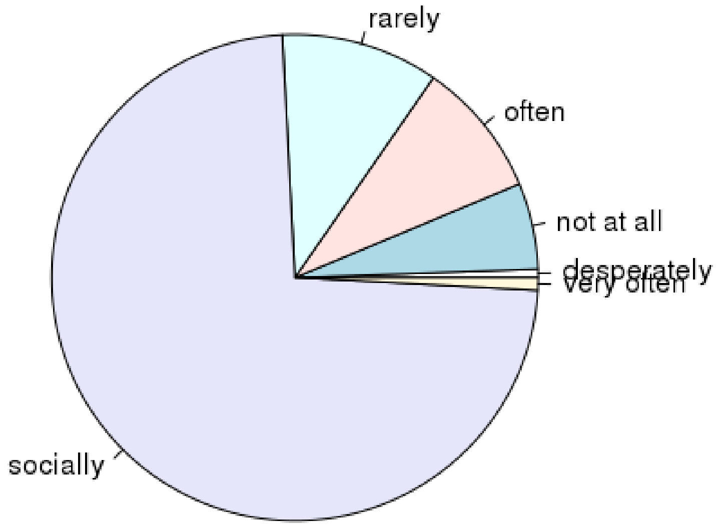
- A: ? barplot
- A: xlab and ylab!

```
> barplot(drinks_table,
        ylab = "Count",
        xlab = "Type of drinker",
        main = "Counts of different types of drinkers")
```

Pie charts

We can also use the pie() function to create pie charts

```
> pie(drinks_table)
```



Questions?

Visualizing Quantitative Data

Visualizing quantitative data: histograms

The first few okcupid users' heights

> profiles$height[1:5]

75, 70, 68, 76, …

To create a histogram we create a set of intervals

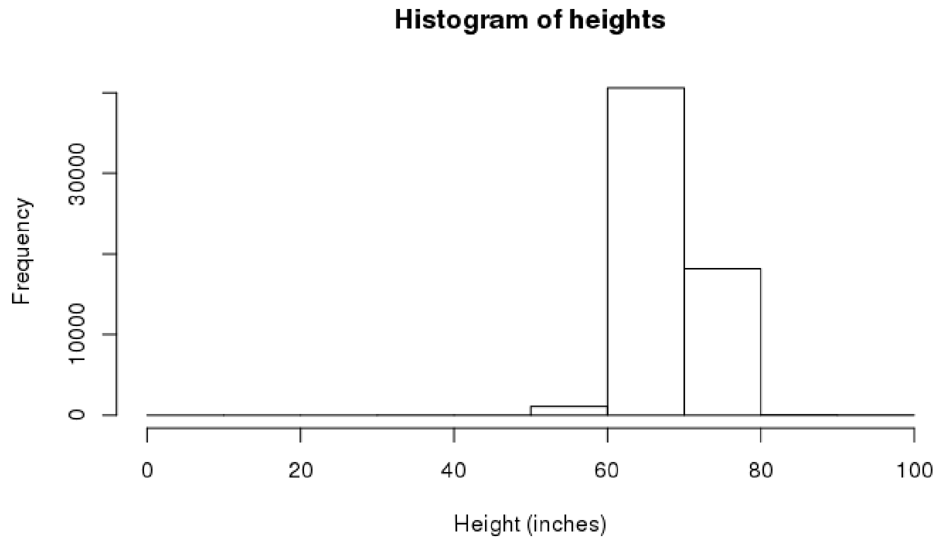
- 60-65, 65-70, 70-75, 75-80

We count the number of points that fall in each interval

We create a bar chart with the counts in each bin

Histograms of heights

Height (inches)	Frequency Count
(0-10]	6
(10-20]	0
(20-30]	1
(30-40]	13
(40-50]	9
(50-60]	1097
(60-70]	40575
(70-80]	18164
(80-90]	50
>90	28



Visualizing heights

We can create histograms in R using the hist() function

Can you create a histogram of heights?

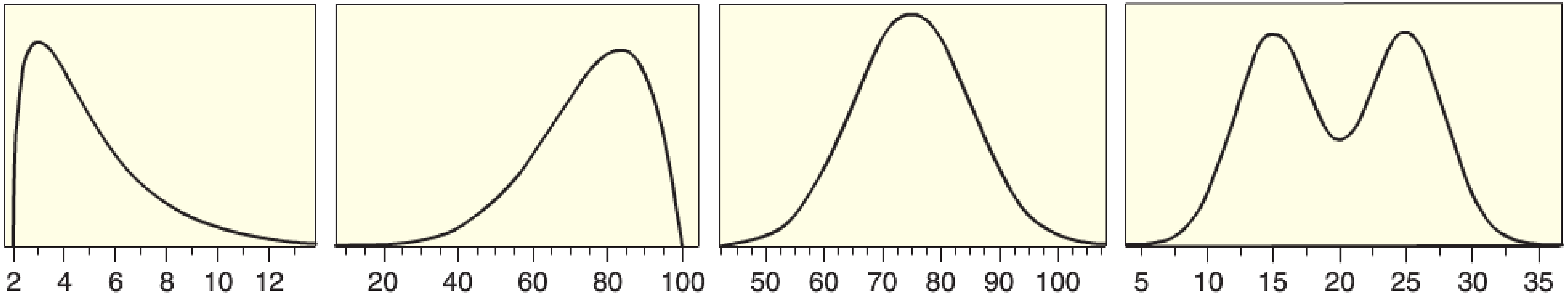
```
> hist(profiles$height)
```

How can you add more bars to the histogram?

- How can we figure out how to add more bars to a histogram?

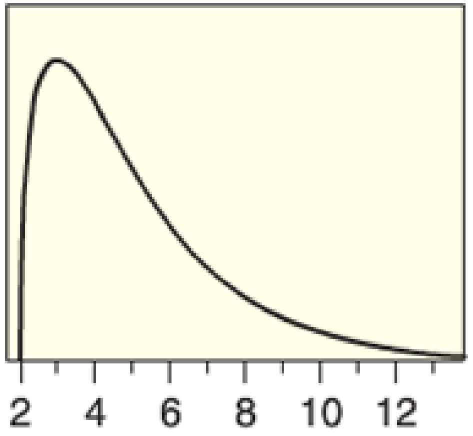
```
> hist(profiles$height, breaks = 50)
```

Common shapes for distributions



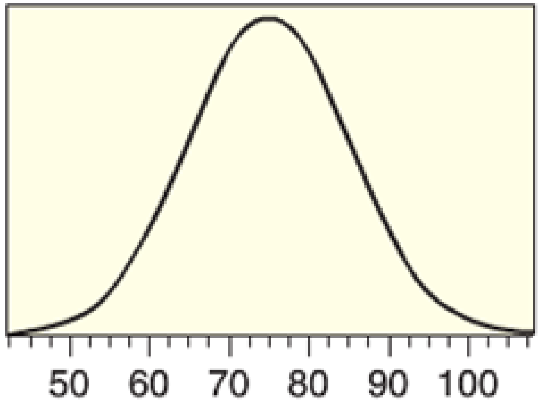
(a) Skewed to the right (b) Skewed to the left (c) Symmetric and bell-shaped (d) Symmetric but not bell-shaped

Can you think of a distribution that is right skewed?



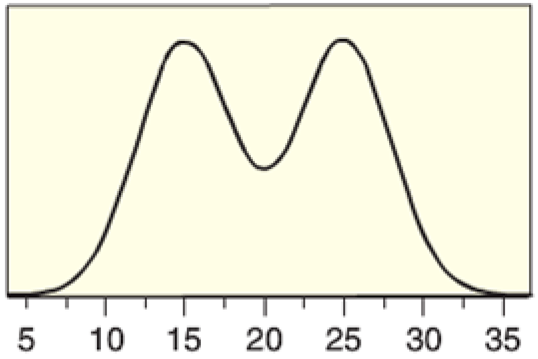
(a) Skewed to the right

Can you think of a distribution that is symmetric and bell-shaped?



(c) Symmetric and bell-shaped

Can you think of a distribution that is symmetric but not bell-shaped?



(d) Symmetric but not bell-shaped

Statistics for quantitative data

Measure of central tendency: The mean

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

```
R: mean(x)
R: mean(x, na.rm = TRUE)
```

The median

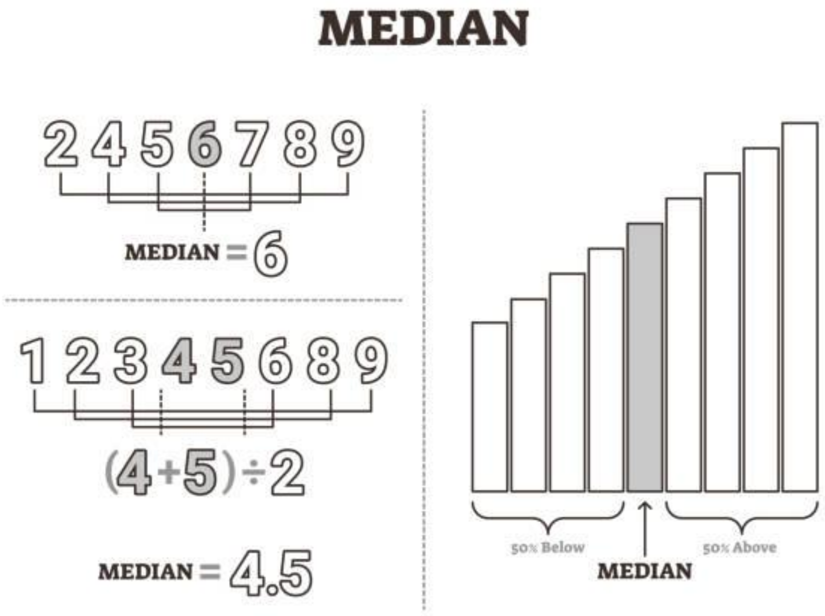
The **median** is the value that splits the data in half

- i.e., half the values are less than the median and half are greater than the median

The median of a data set of size n is:

- If n is odd: The middle value of the sorted data

- If n is even: The average of the middle two values of the sorted data




```
R: median(v)
   median(v, na.rm = TRUE)
```

Resistance

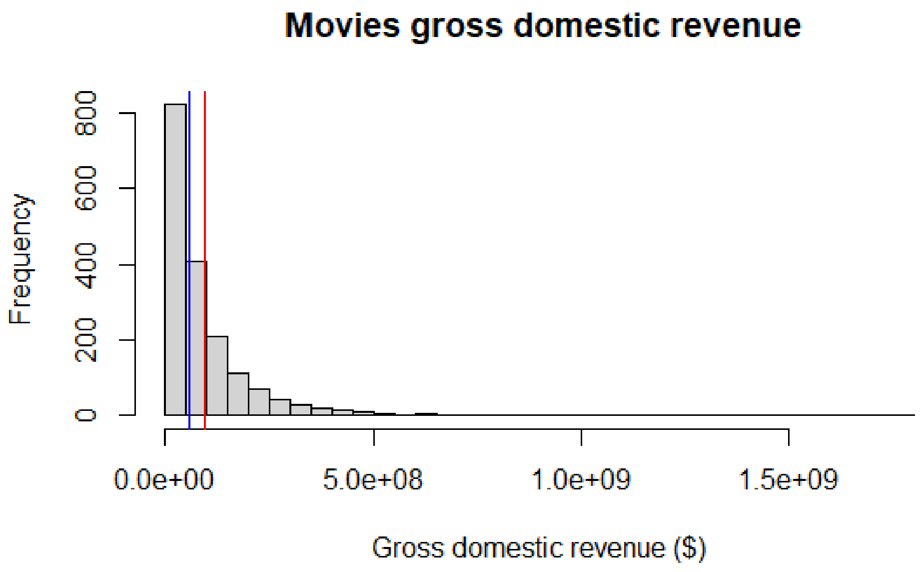
We say that a statistics is **resistant** if it is relatively unaffected by extreme values (outliers)

The median is resistant when the mean is not

Example:

Mean US salary = $72,641

Median US salary = $51,939



Let’s try it in RStudio!

Measures of spread

Measure of spread 1: The standard deviation

The standard deviation can be computed using the following formula:

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Example: computing the standard deviation

Suppose we had a sample with n = 4 points:

$x_1 = 8$, $x_2 = 2$, $x_3 = 6$, $x_4 = 4$,

We can compute the mean using the formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \tfrac{1}{4} \cdot (x_1 + x_2 + x_3 + x_4) \quad = \tfrac{1}{4} \cdot (8 + 2 + 6 + 4) \;= 5$$

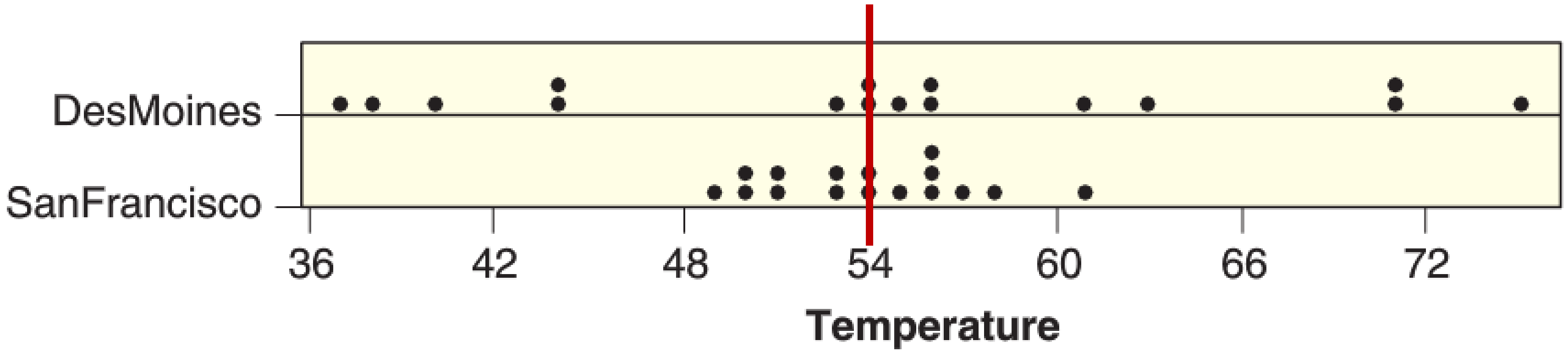
The standard deviation can be computed using the formula:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(remember order of operations!)

Average monthly temperature:
Des Moines vs. San Francisco

Data measured on April 14th from 1997 to 2010:



Mean temperature (°F): Des Moines = 54.49 San Francisco = 54.01

Standard deviation (°F): Des Moines = 11.73 San Francisco = 3.38

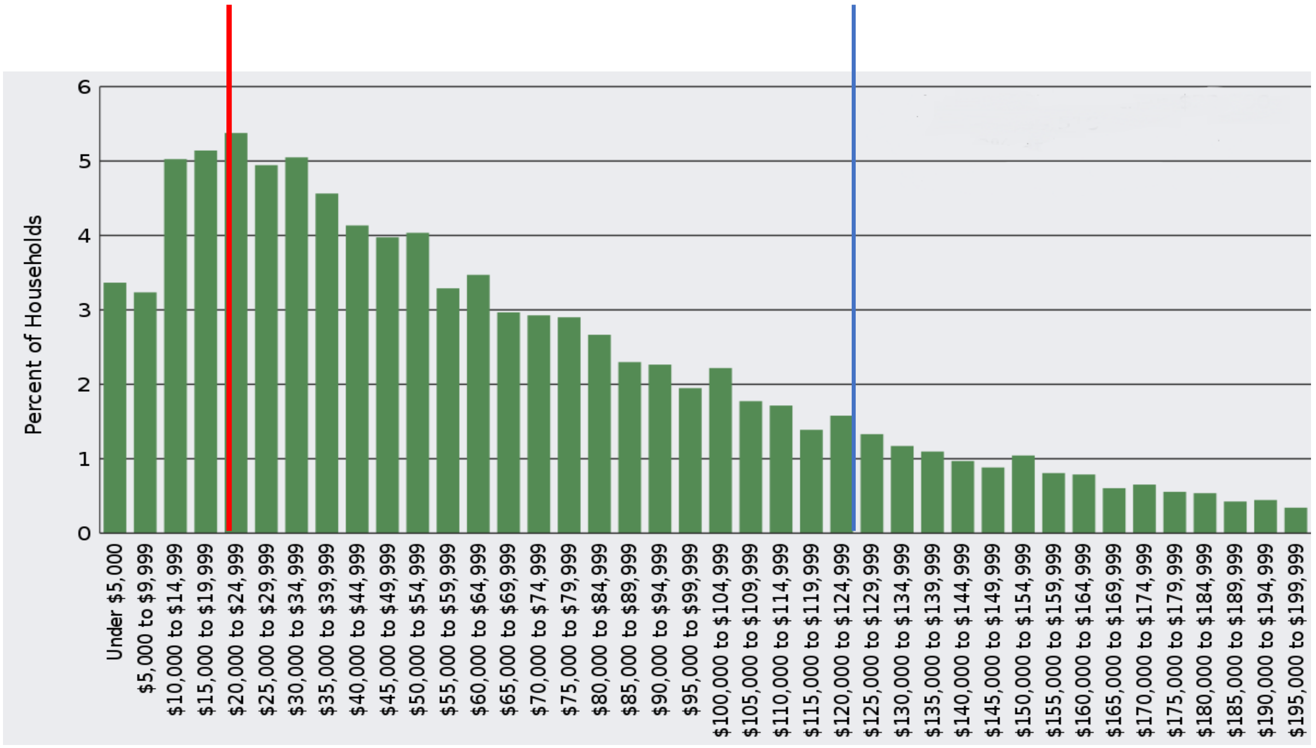
Percentiles

The **Pth percentile** is the value of a quantitative variable which is greater than P percent of the data

For the US income distribution what are the 20th and 80th percentiles?

20th percentile = $21,430

80th percentile = $112, 254



R: `quantile(v, .95)`

Five Number Summary

A **five-number summary** is a set of five descriptive statistics that provides a concise overview of a dataset's distribution.

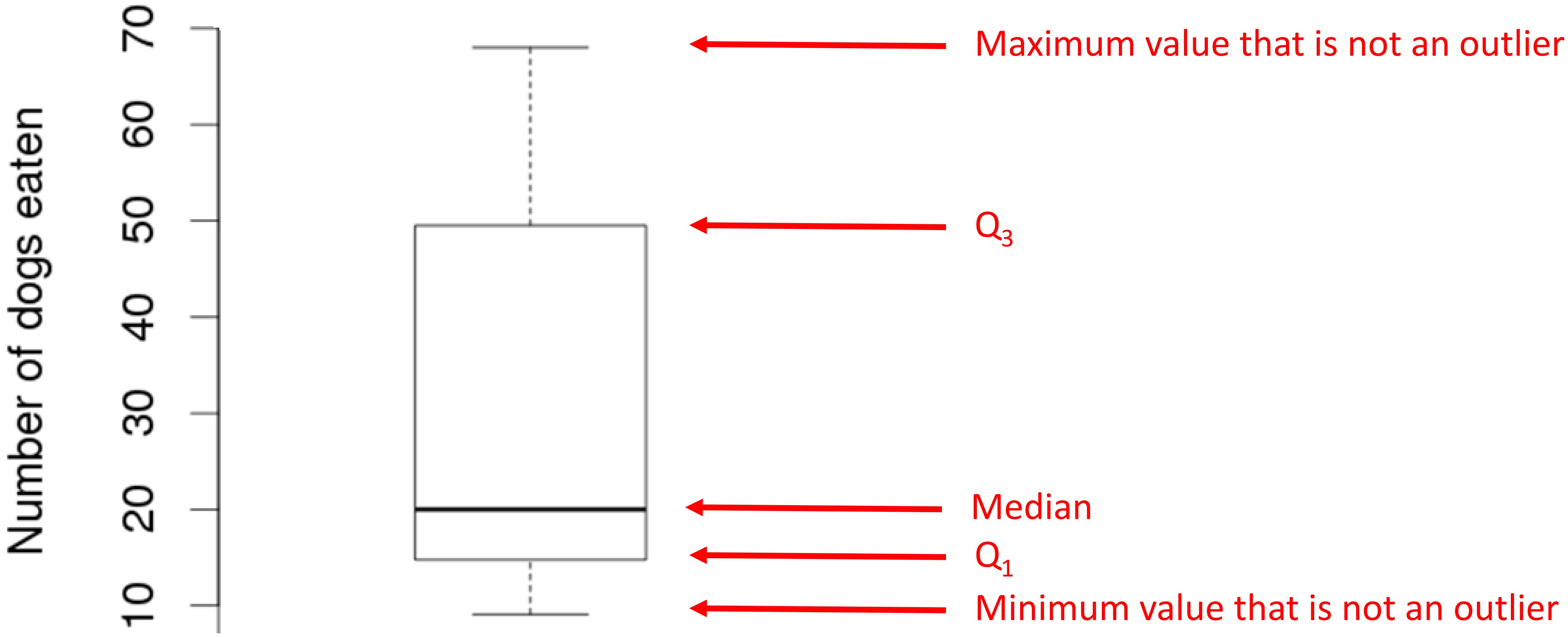
Five Number Summary = (minimum, Q_1, median, Q_3, maximum)

Q_1 = 25th percentile (also called 1st quartile)
Q_3 = 75th percentile (also called 3rd quartile)

Roughly divides the data into fourths

Measure of spread 2: **Interquartile range (IQR)** = $Q_3 - Q_1$

Box plots can also visualize quantitative data

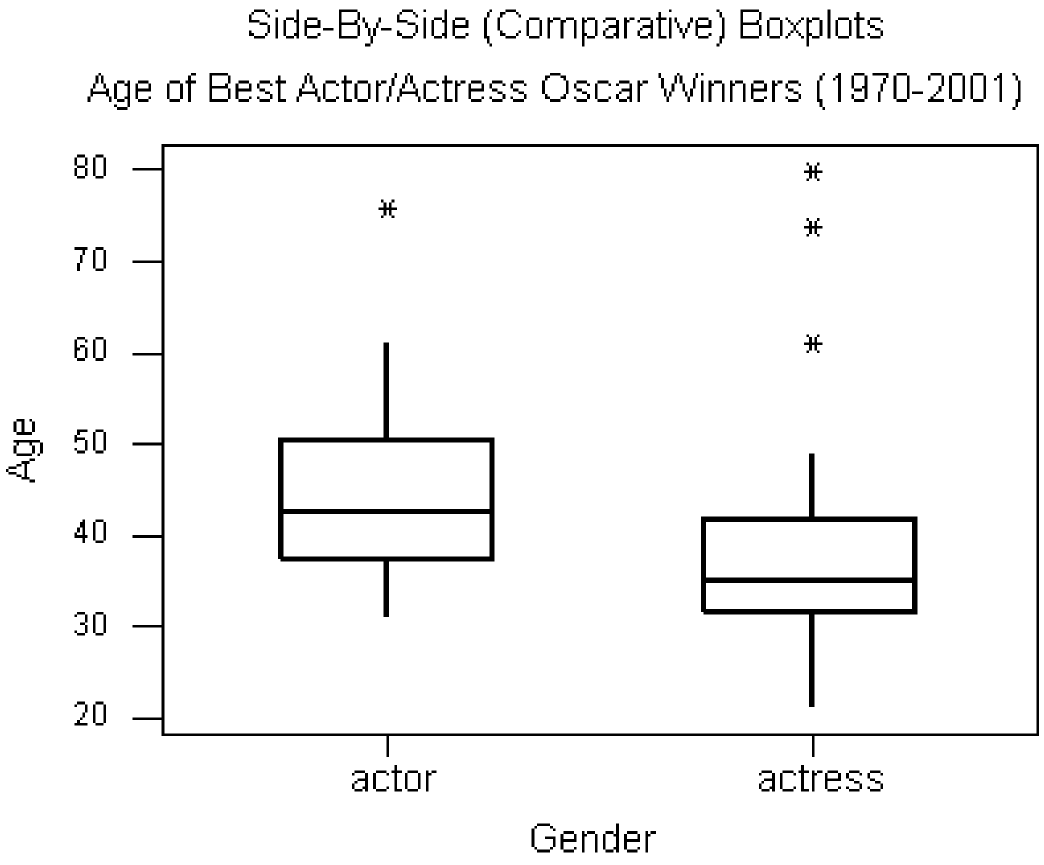


R: `boxplot(v)`

Side-by-side boxplots

Boxplots are particularly useful for comparing distributions!

What does the figure above show?



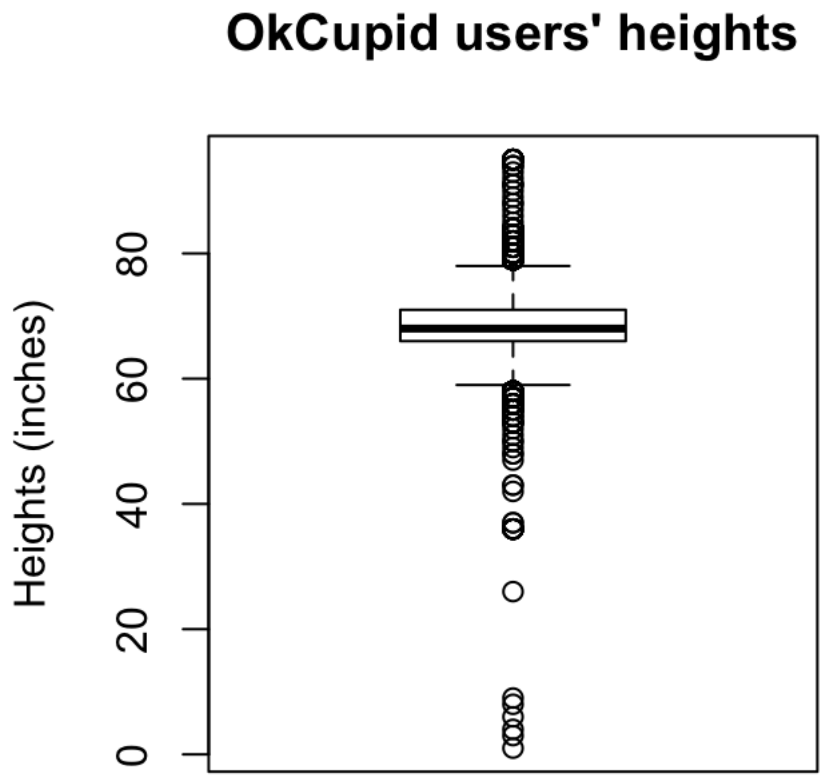
Outliers

Outliers on boxplots are values that are more than 1.5 * IQR

What should we do if we have outliers?

Investigate!

- If there are due to an error, remove them



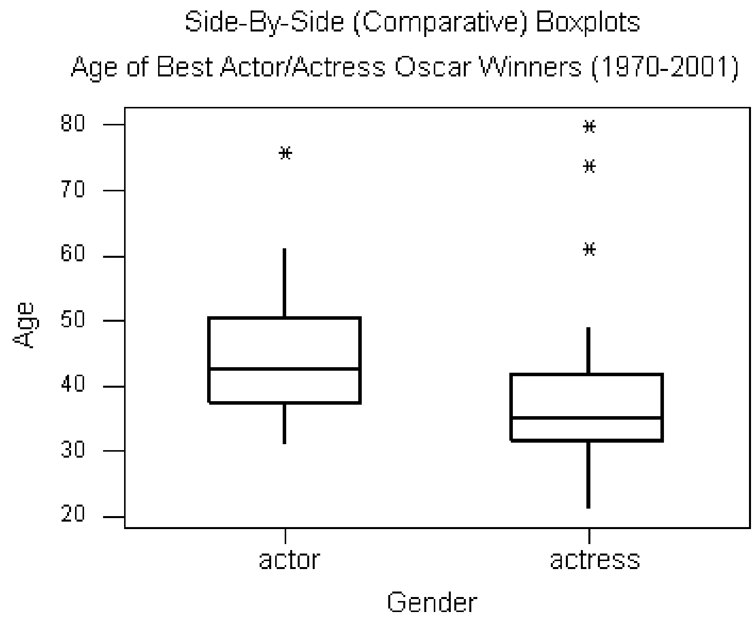
Outliers

Outliers on boxplots are values that are more than 1.5 * IQR

What should we do if we have outliers?

Investigate:

- If there are due to an error, remove them
- **If not, need to account for them**



Questions?

Let’s try it in RStudio!

Visualizing two quantitative variables

CitiBike data

Let’s look at the bike share data from NYC

> load('daily_bike_totals.rda’)

[CitiBike analysis]

What does each case correspond to?

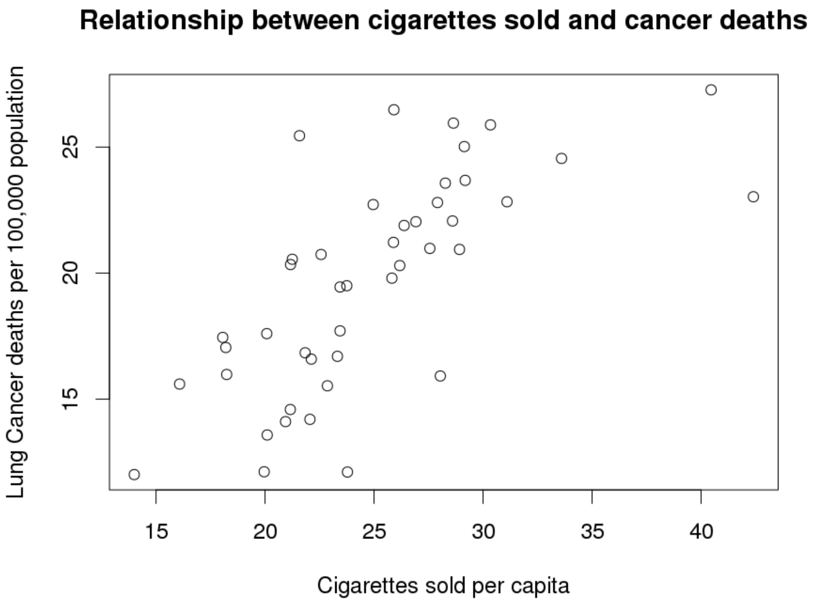
We can use the dim() function to get how many cases and variables there are

- How many are there?

Scatterplot

A **scatterplot** graphs the relationship between two variables

- Each axis represents the value of one variables
- Each point the plot shows the value for the two variables for a single data case



If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis.

Scatter plots

We can use the plot(x, y) function to create scatter plots

Can you create a scatter plot of the relationship between the minimum and maximum temperatures?

```
> plot(bike_daily_data$min_temperature,
       bike_daily_data$max_temperature,
       xlab = "Minimum temperature",
       ylab = "Maximum temperature",
       main = "Relationship between min and temp")
```

Plotting time series

We can use the plot(x, y) function to plot time series

```
# we can connect the points in a plot using
> plot(x, y, type = 'l')    # connected points
> plot(x, y, type = 'o')   # both points and dots

> plot(bike_daily_data$date,  bike_daily_data$trips,
          type = 'o',
          xlab = "Date",
          ylab = "Number of trips",
          main = "Total number of trips on each day")
```

The correlation coefficient

The **correlation** is measure of the strength and direction of a linear association between two variables

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

R: `cor(x, y)`

Properties of the correlation

Correlation as always between -1 and 1: $-1 \leq r \leq 1$

The sign of r indicates the direction of the association

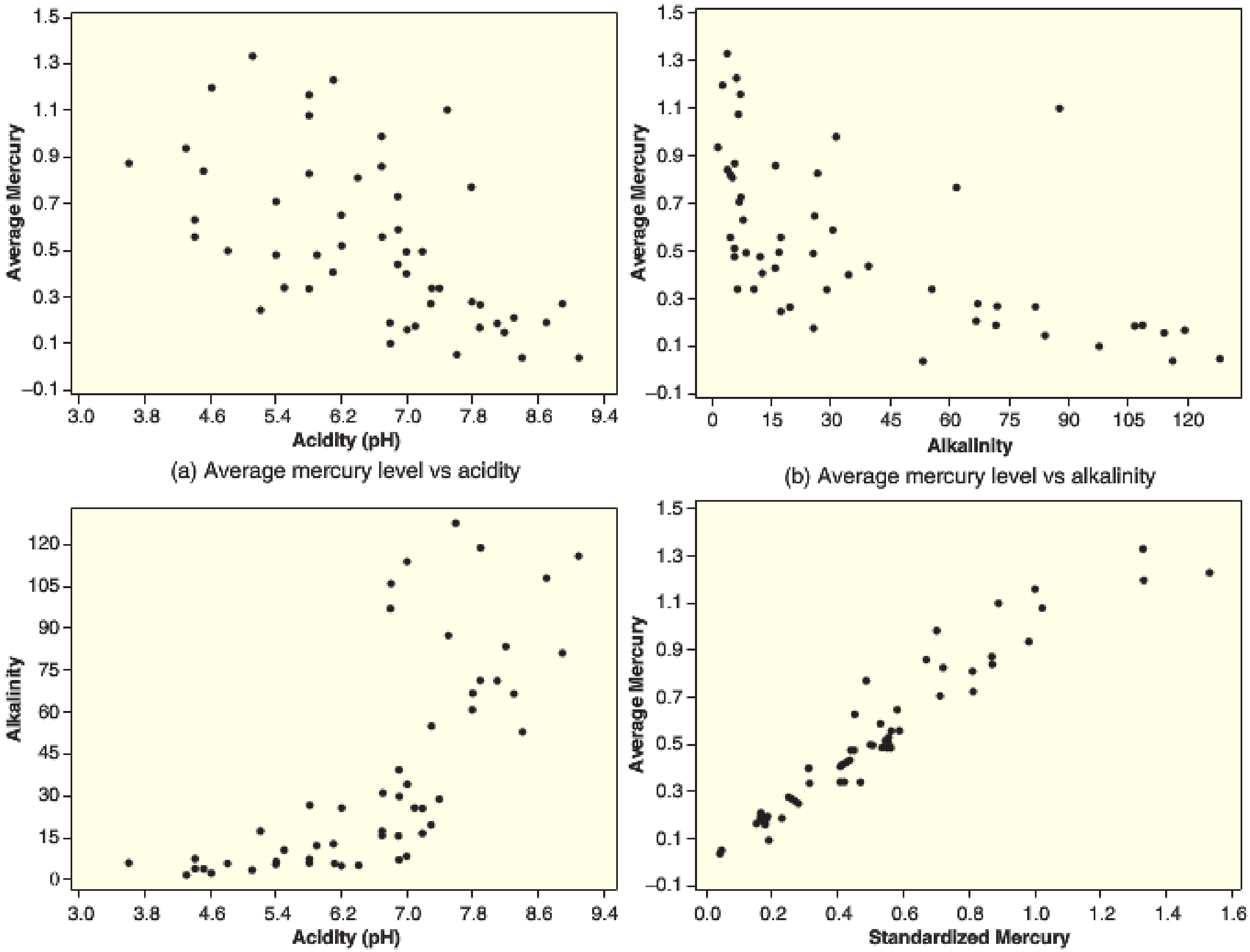
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Correlation is symmetric: r = cor(x, y) = cor(y, x)

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

Florida lakes

[Correlation game]



(a) Average mercury level vs acidity

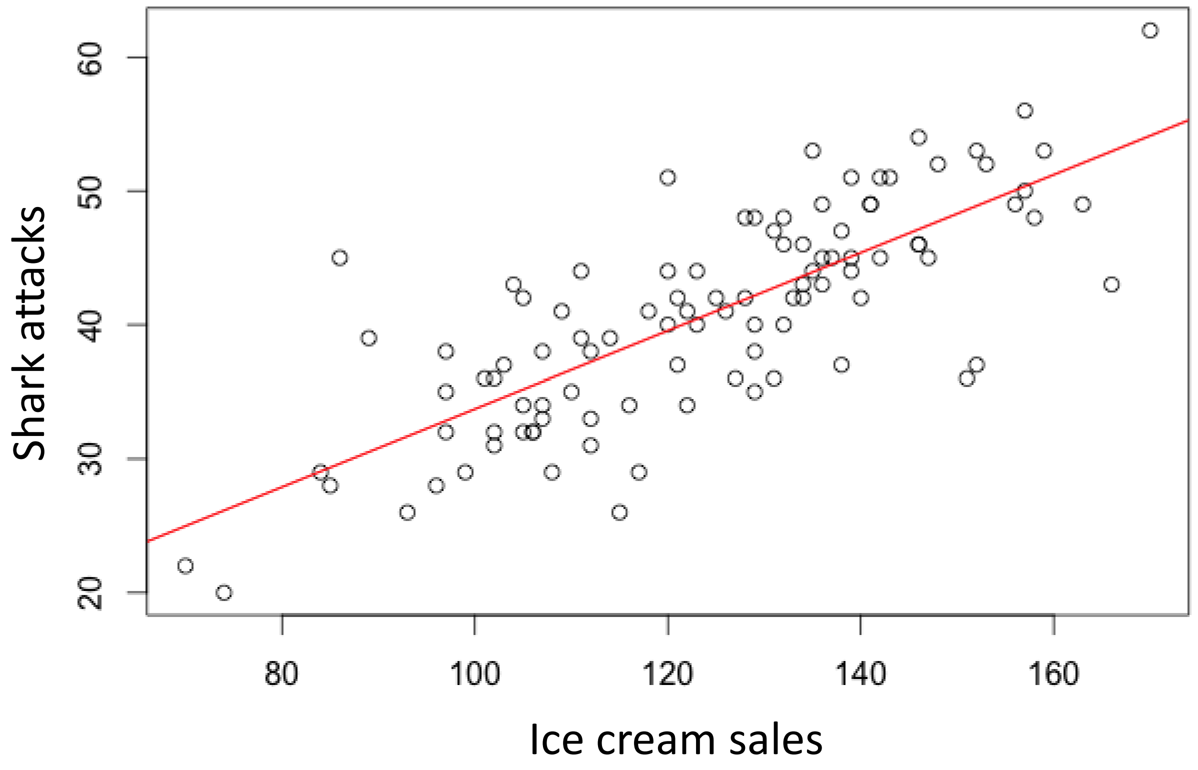
(b) Average mercury level vs alkalinity

(c) Alkalinity vs acidity

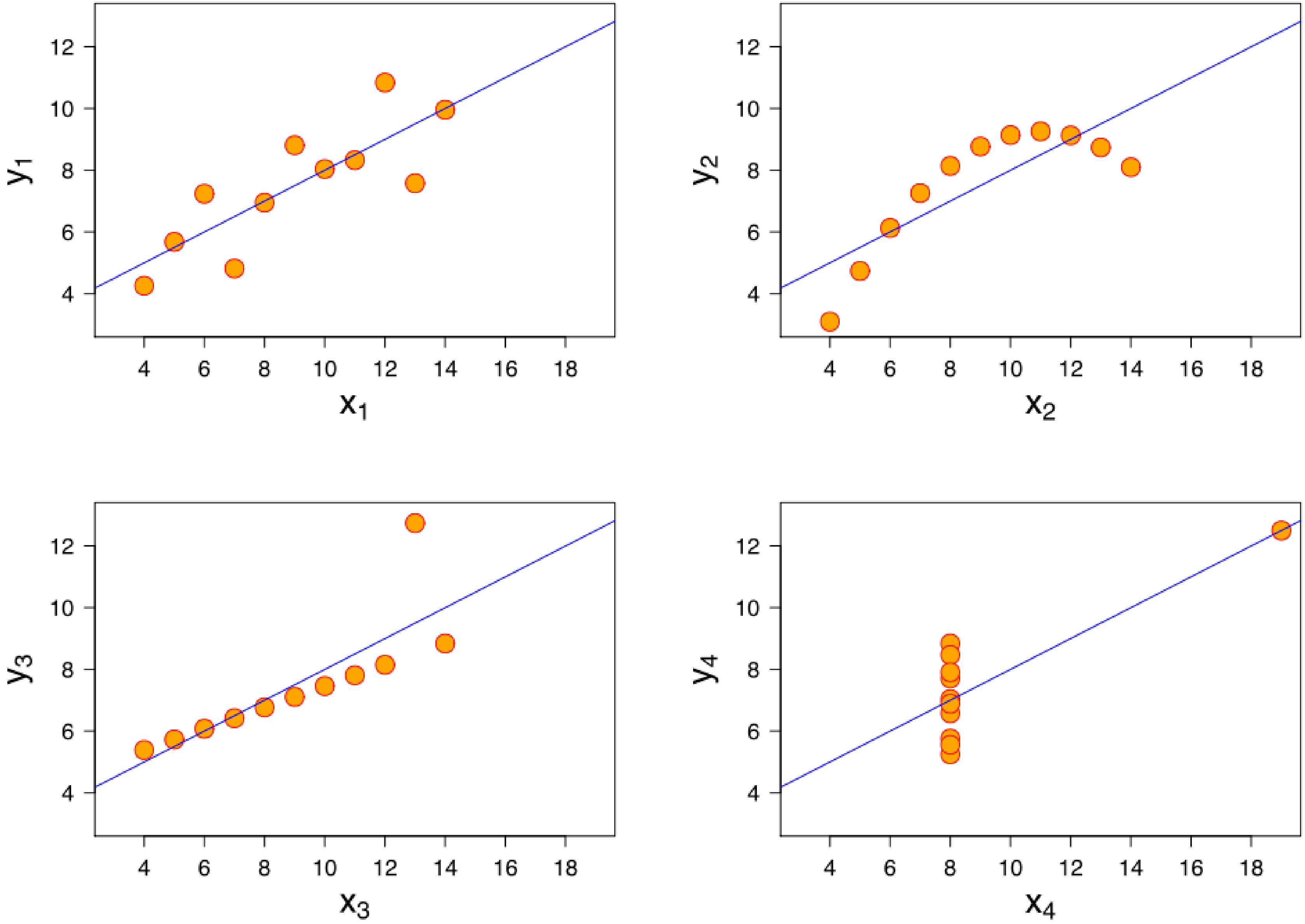
(d) Average vs standardized mercury levels

Correlation caution #1

A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



Anscombe's quartet (r = 0.81)



Next class: data transformations with dplyr