Chapter 1 Section 2: Data basics

Some definitions

- **Data:**

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:**

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.
- **Data matrix:**

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.
- **Data matrix:** A collection of data with each row a case and each column a variable.

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.
- **Data matrix:** A collection of data with each row a case and each column a variable.
- **Case:**

Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.
- **Data matrix:** A collection of data with each row a case and each column a variable.
- **Case:** An observational unit.

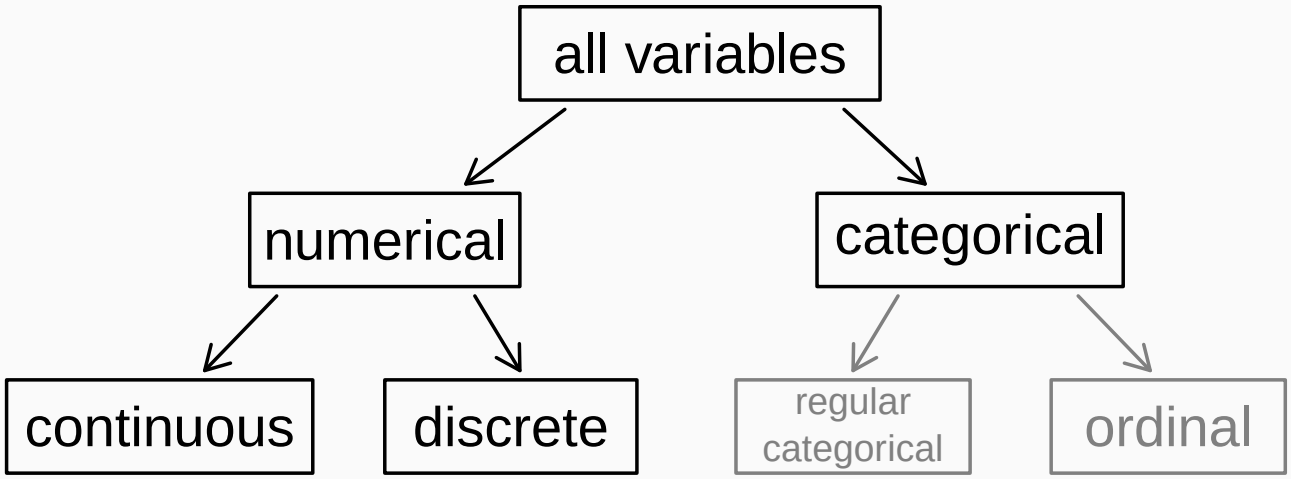
Some definitions

- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.
- **Data matrix:** A collection of data with each row a case and each column a variable.
- **Case:** An observational unit.
- **Variable:**

Some definitions

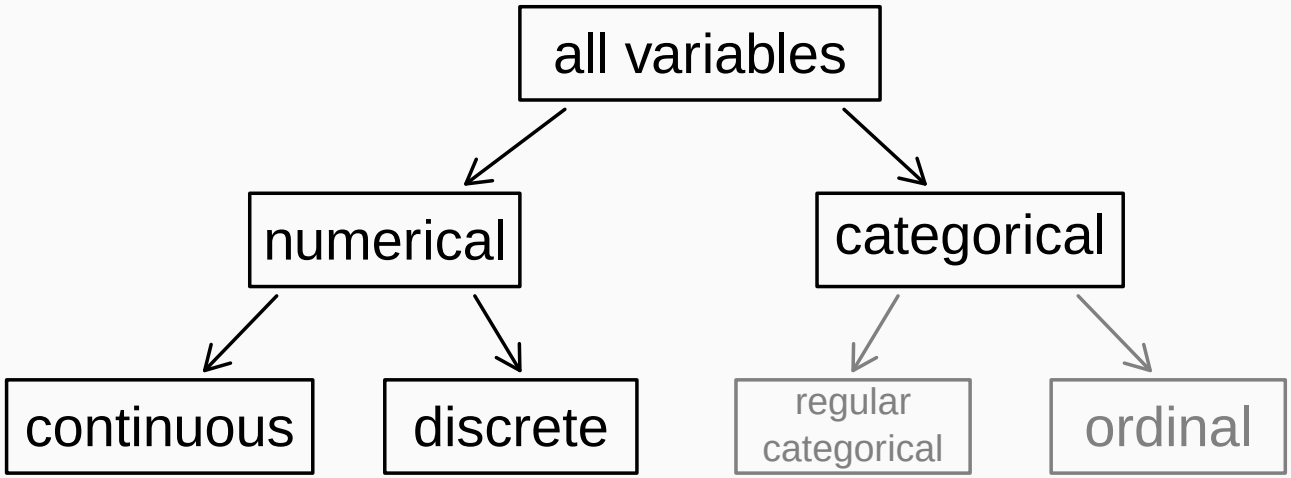
- **Data:** Observations, measurements, and information that is analyzed.
- **Summary statistic:** A single number summarizing a large amount of data.
- **Data matrix:** A collection of data with each row a case and each column a variable.
- **Case:** An observational unit.
- **Variable:** A characteristic (usually one of many) that is measured from each case.

Types of variables



- Numerical variables take values that can be added, subtracted, and averaged in a sensible way.
- Discrete numerical variables take on values with jumps e.g. counts, “how many”.
- Continuous numerical variables take on values without jumps e.g. weights, heights, ”how much”.

Types of variables 2



- Categorical variables do not take values that can be added, subtracted, and averaged in a sensible way.

Practice

```
> mtcars
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
```

Cases? Variables? Types of Variables?

Variable descriptions

```
mtcars
A data frame with 32 observations on 11 variables.

[, 1]  mpg   Miles/(US) gallon
[, 2]  cyl   Number of cylinders
[, 3]  disp  Displacement (cu.in.)
[, 4]  hp    Gross horsepower
[, 5]  drat  Rear axle ratio
[, 6]  wt    Weight (1000 lbs)
[, 7]  qsec  1/4 mile time
[, 8]  vs    V/S
[, 9]  am    Transmission (0 = automatic, 1 = manual)
[,10]  gear  Number of forward gears
[,11]  carb  Number of carburetors
```

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`:

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`:

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*
- `bedtime`:

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*
- `bedtime`: *categorical, ordinal*

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*
- `bedtime`: *categorical, ordinal*
- `countries`:

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*
- `bedtime`: *categorical, ordinal*
- `countries`: *numerical, discrete*

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*
- `bedtime`: *categorical, ordinal*
- `countries`: *numerical, discrete*
- `dread`:

Types of variables (cont.)

	gender	sleep (hr)	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- `gender`: *categorical*
- `sleep`: *numerical, continuous*
- `bedtime`: *categorical, ordinal*
- `countries`: *numerical, discrete*
- `dread`: *categorical, ordinal - could also be used as numerical*

Practice

What type of variable is a telephone area code?

(a) numerical, continuous

(b) numerical, discrete

(c) categorical

(d) categorical, ordinal

Practice

What type of variable is a telephone area code?

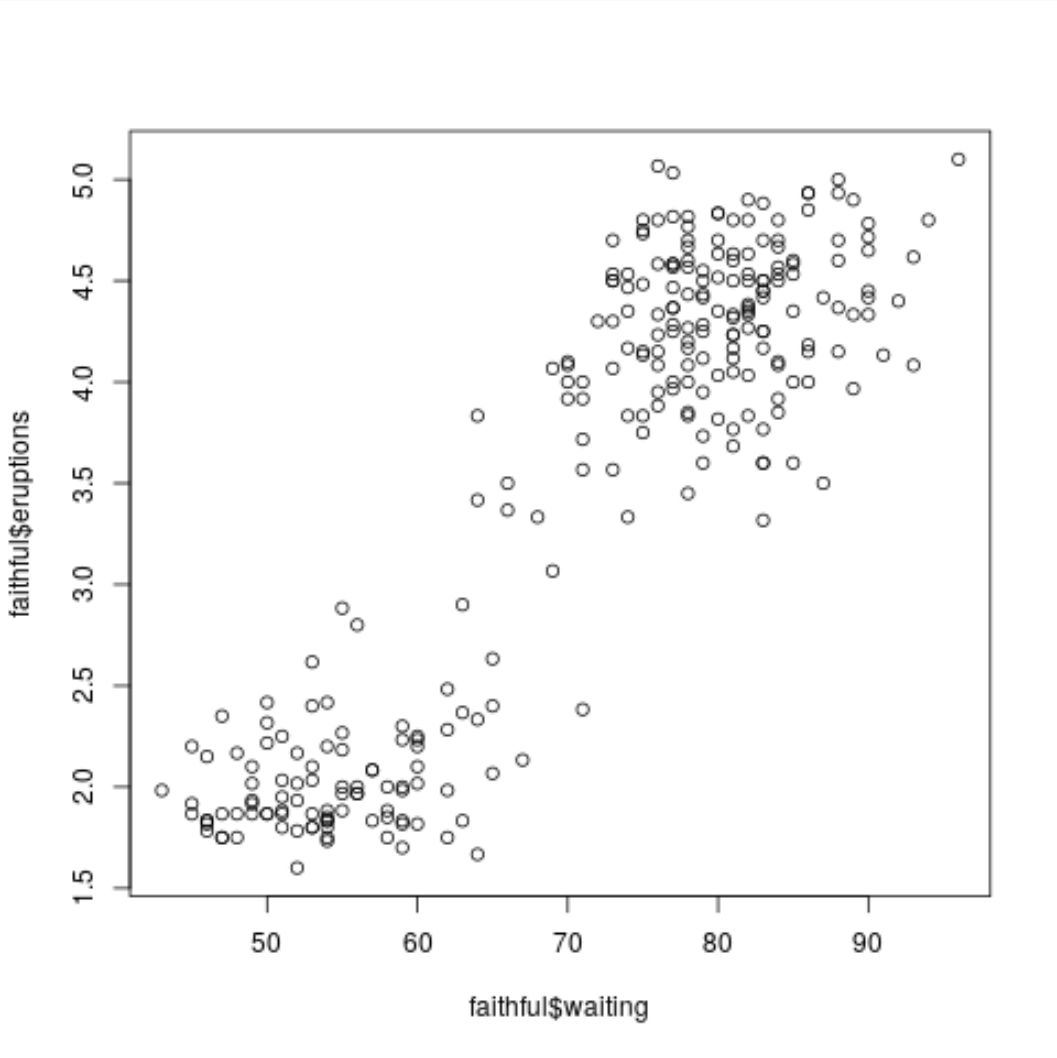
(a) numerical, continuous

(b) numerical, discrete

(c) *categorical*

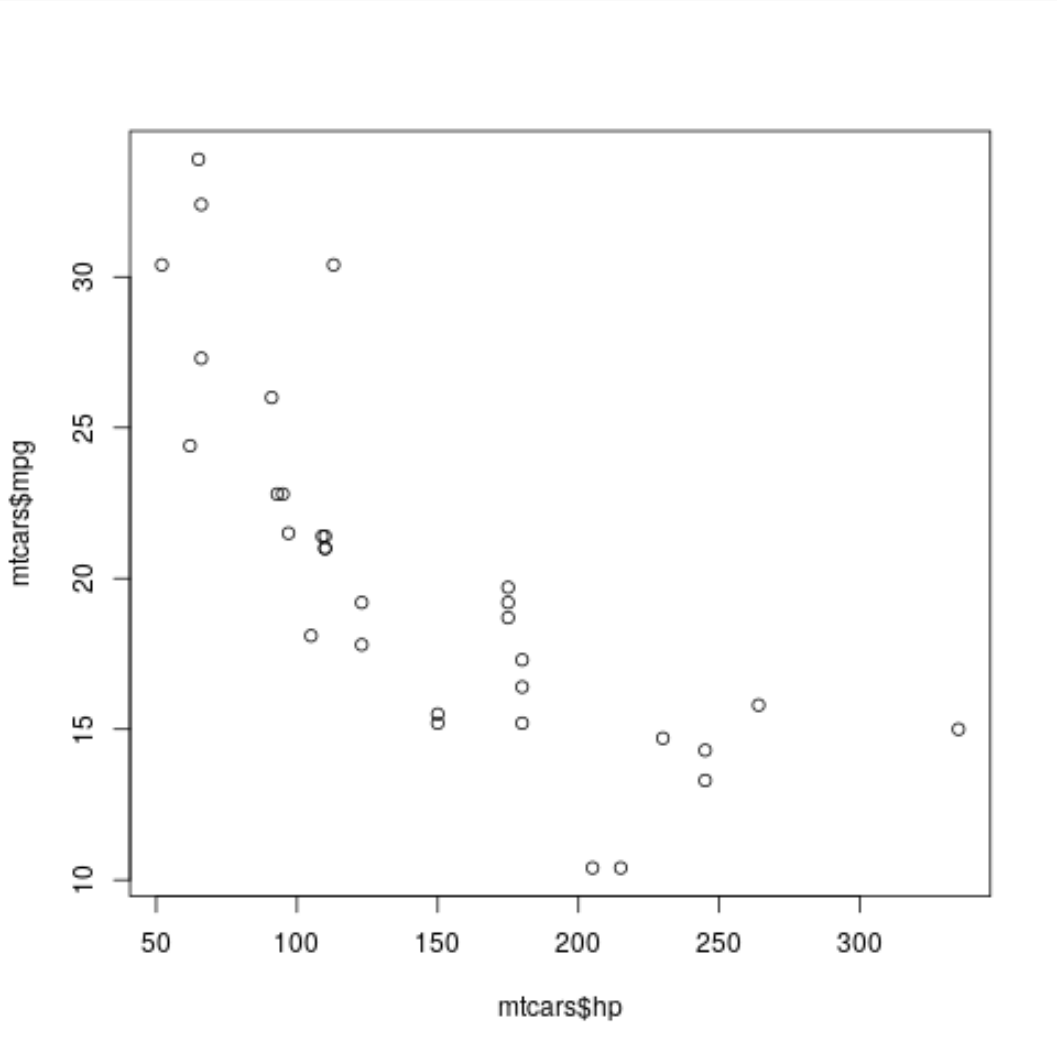
(d) categorical, ordinal

Positive association



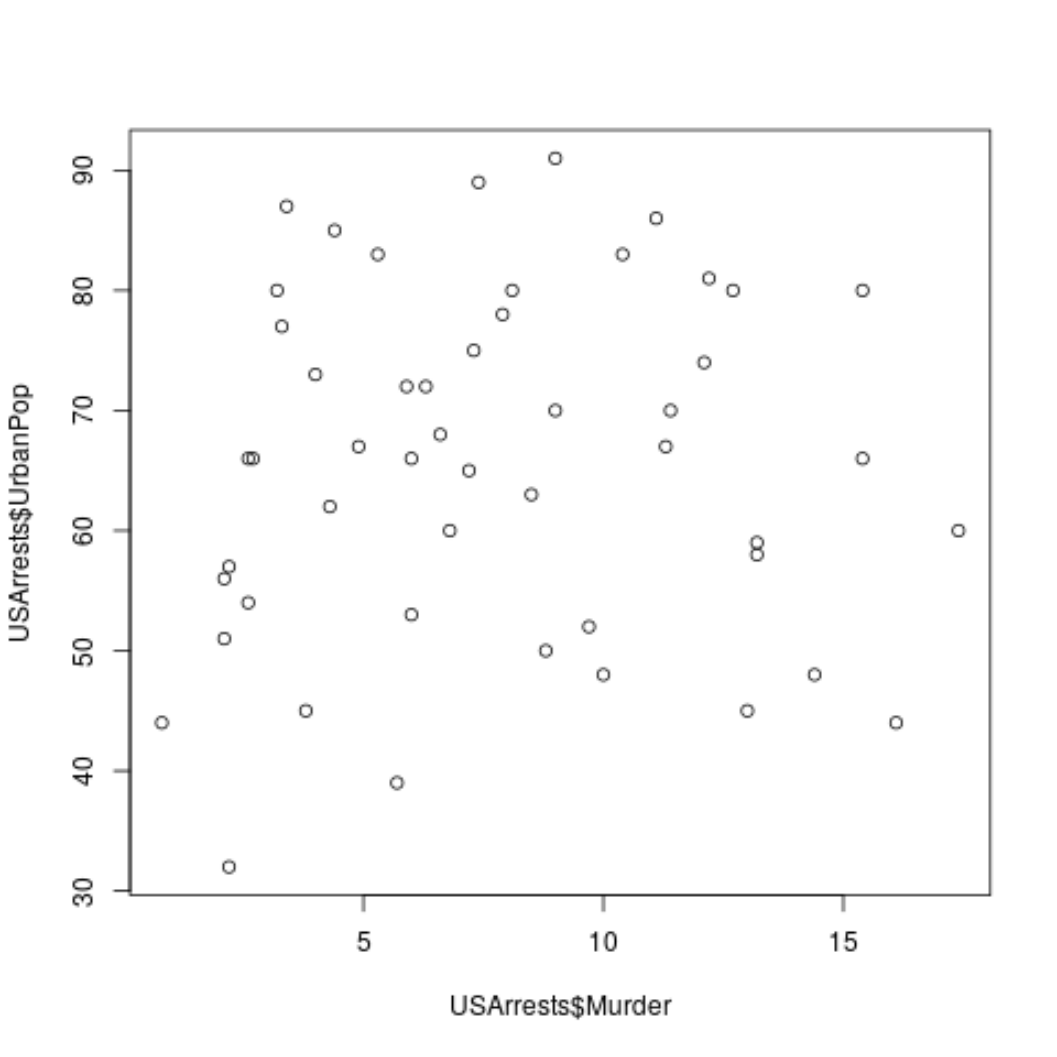
The eruption time (min) vs wait time (min) for 272 cases of Old Faithful.

Negative association



The mpg vs. HP for 32 cars.

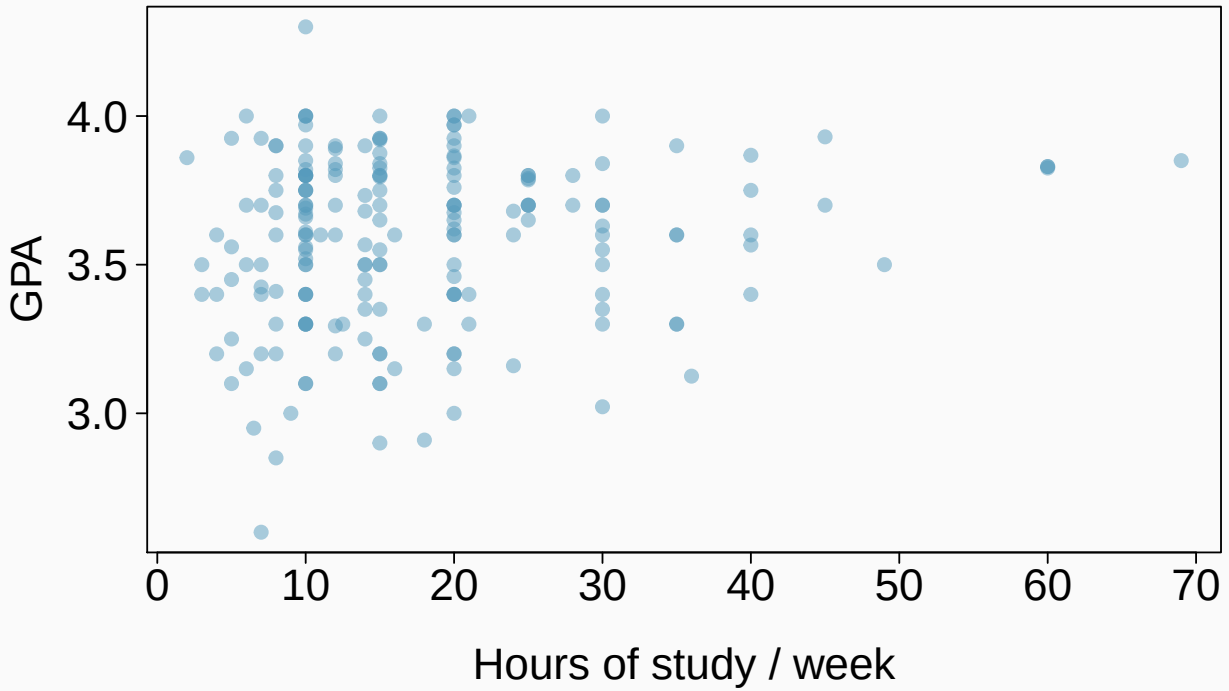
Independent variables



From 1973, murder rate vs urban population proportion (50 states).

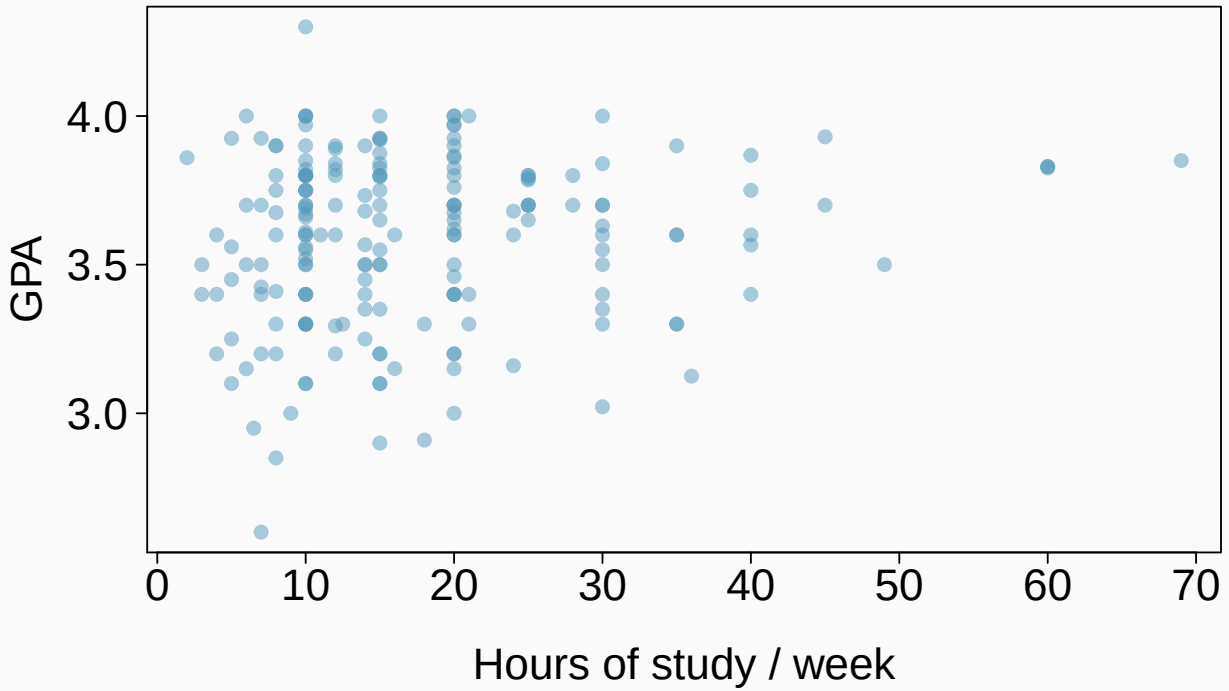
Relationships among variables

Does there appear to be a relationship between GPA and number of hours students study per week?



Relationships among variables

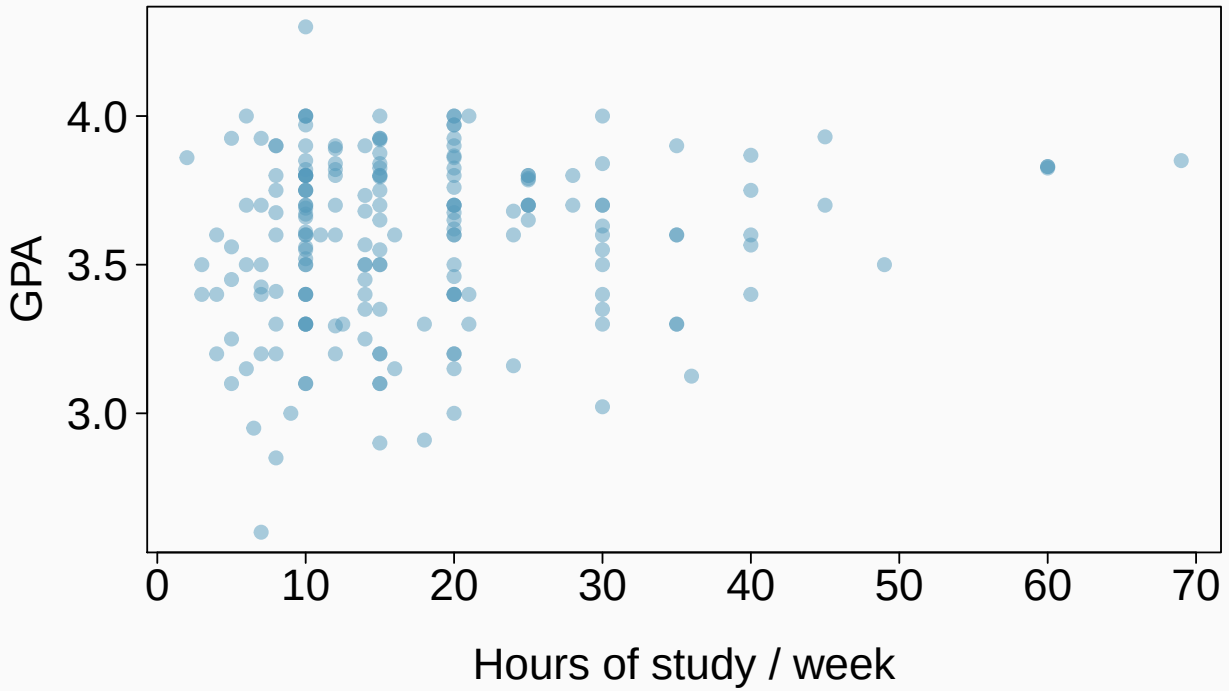
Does there appear to be a relationship between GPA and number of hours students study per week?



Can you spot anything unusual about any of the data points?

Relationships among variables

Does there appear to be a relationship between GPA and number of hours students study per week?

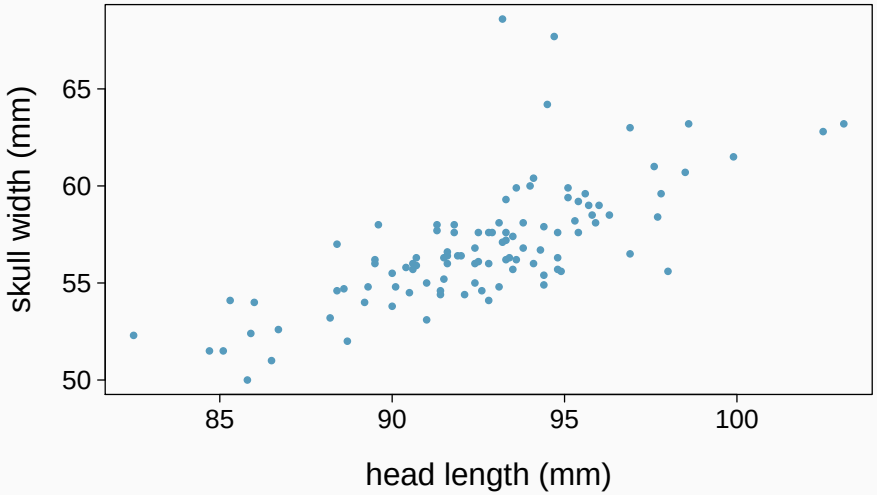


Can you spot anything unusual about any of the data points?

There is one student with GPA > 4.0, this is likely a data error.

Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) Head length and skull width are positively associated.

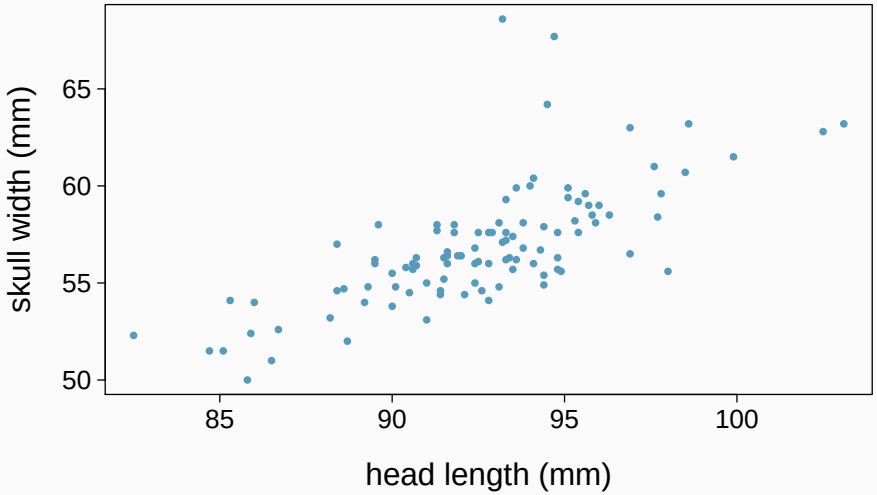
(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) *Head length and skull width are positively associated.*

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Associated vs. independent

- When two variables show some connection with one another, they are called *associated* variables.
  - Associated variables can also be called *dependent* variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.

Class survey

What would be some interesting questions we could ask everyone in the room?

For each question, what type of variable would be recorded?

Would the survey be anonymous?

Which variables would you expect to be associated? independent?