Week 3: Basic regression

1.Relationship between two numerical variables

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Today: Relationship between two numerical variables

2. Main ideas
    1. Scatterplot
    2. Correlation
    3. Make a scatterplot with ggplot

3. Summary

Outline

1. Today: Relationship between two numerical variables

2. Main ideas
    1. Scatterplot
    2. Correlation
    3. Make a scatterplot with ggplot

3. Summary

Outline

1. Today: Relationship between two numerical variables

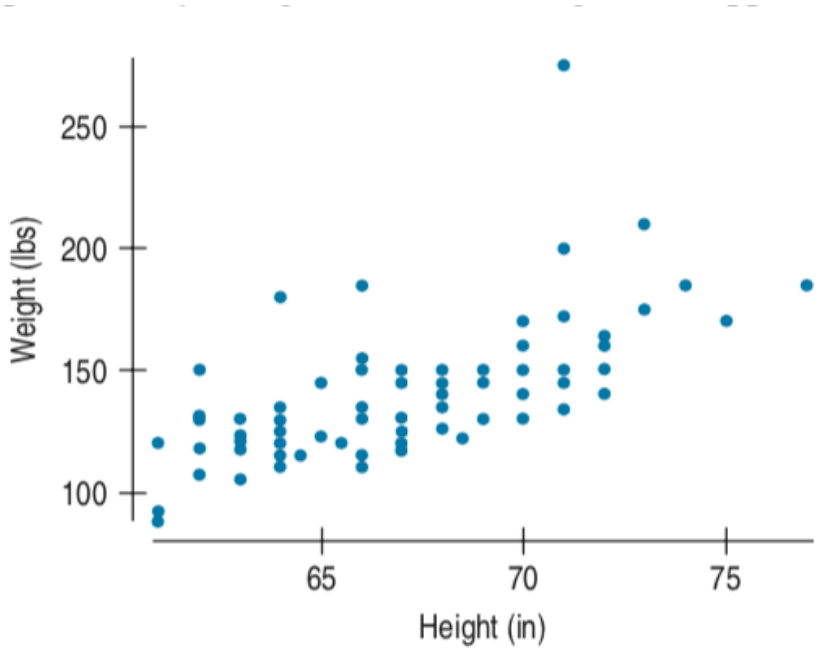
2. Main ideas
    1. Scatterplot
    2. Correlation
    3. Make a scatterplot with ggplot

3. Summary

Scatterplots display the relationship between two numerical variables.

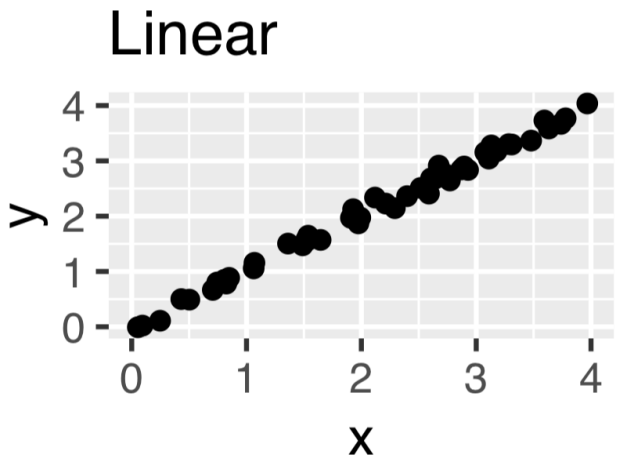
The x axis often displays the **explanatory** or **predictor** variable. The y axis often displays the **response** variable.

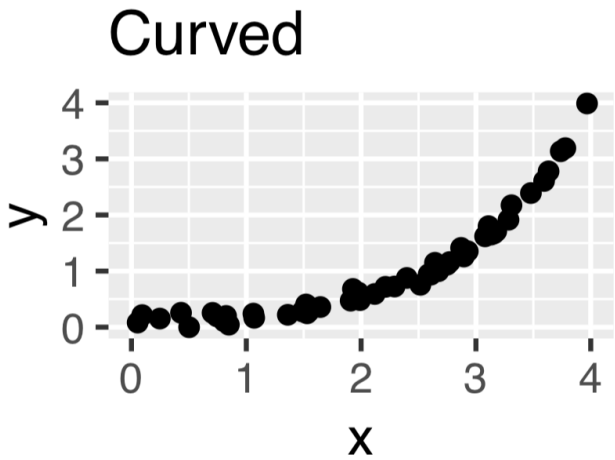
Below is a scatterplot that displays the relationship between the height variable and the weight variable.



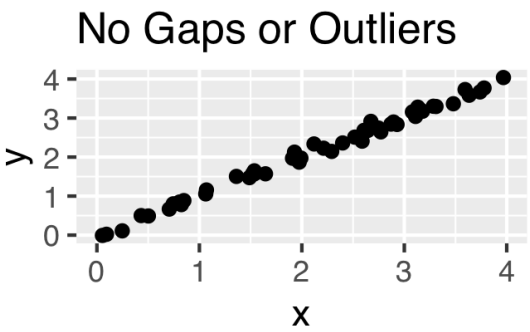
When I look at a scatter plot, I'm evaluating four characteristics of the plot:

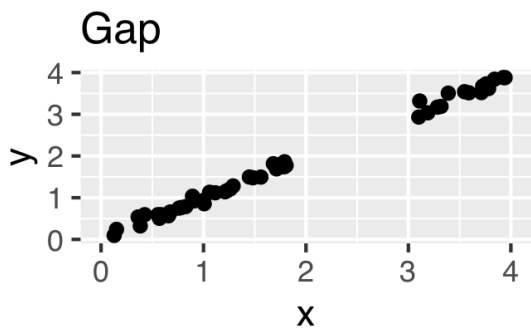
1. Shape (linear or curved)
2. Gaps or outliers
3. Direction (positive or negative association)
4. Strength (weak, moderate, or strong)

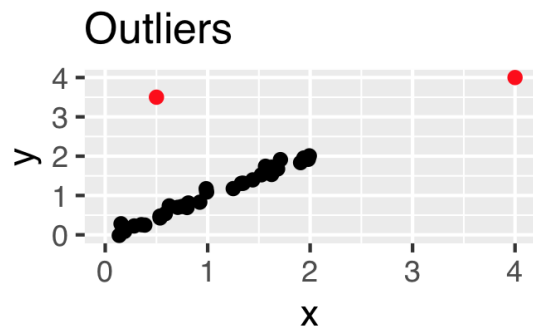

Linear
y
x
0
1
2
3
4


Curved
y
x
0
1
2
3
4

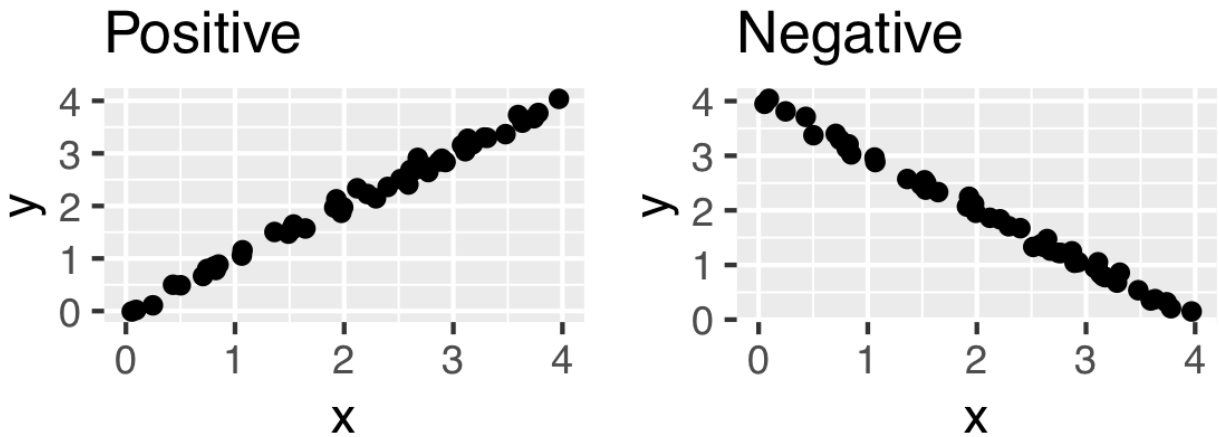
(2) Gaps or outliers



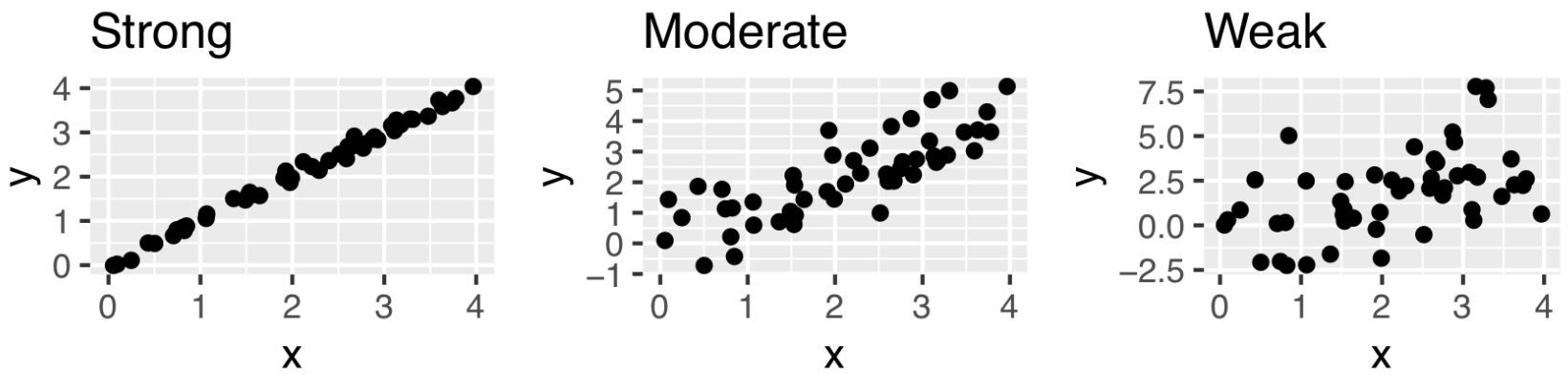




- Positive association: as the x variable increases, the y variable also increases
- Negative association: as the x variable increases, the y variable decreases



The relationship is strong if the points fall close to the trend, and weak if they do not.



1. Today: Relationship between two numerical variables

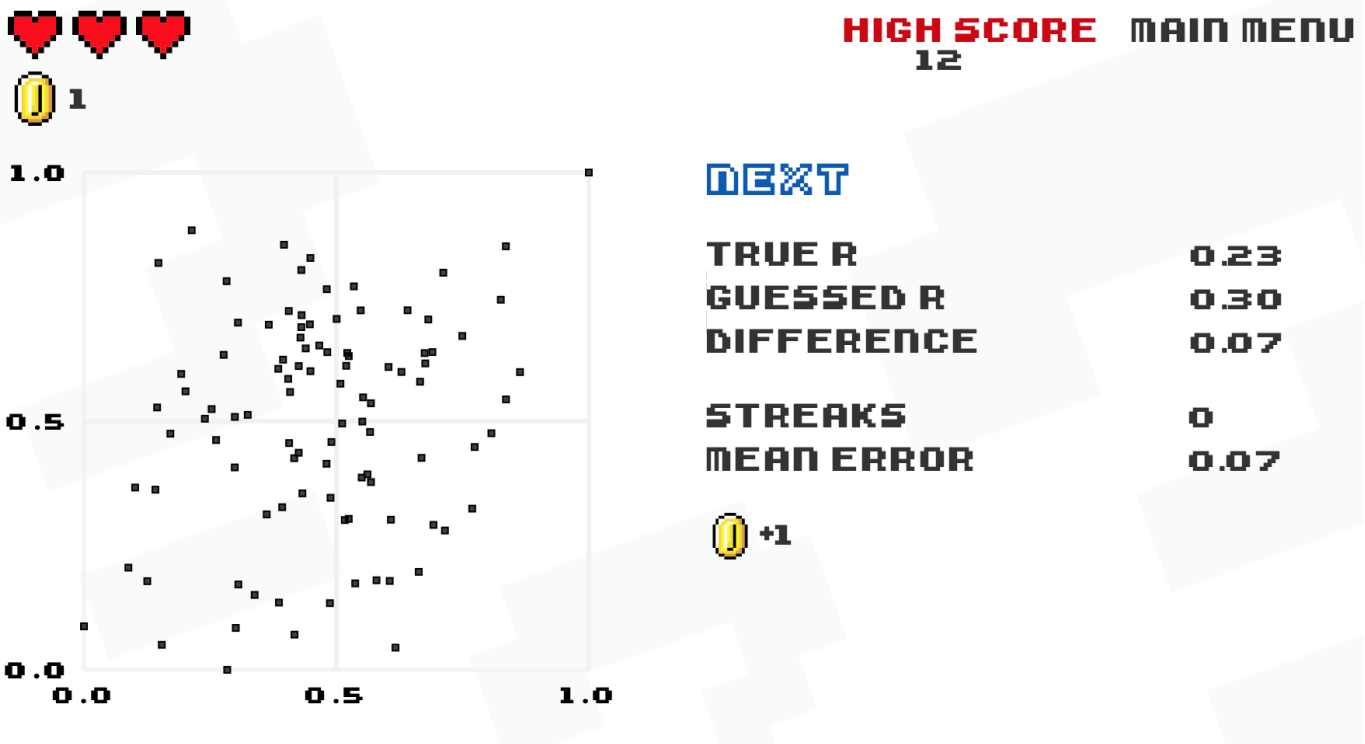
2. Main ideas
    1. Scatterplot
    2. Correlation
    3. Make a scatterplot with ggplot

3. Summary

Given a visualization of the relationship between two numerical variables, one needs to find ways to summarize the information that facilitate understanding and insight. One standard tool is **correlation**.

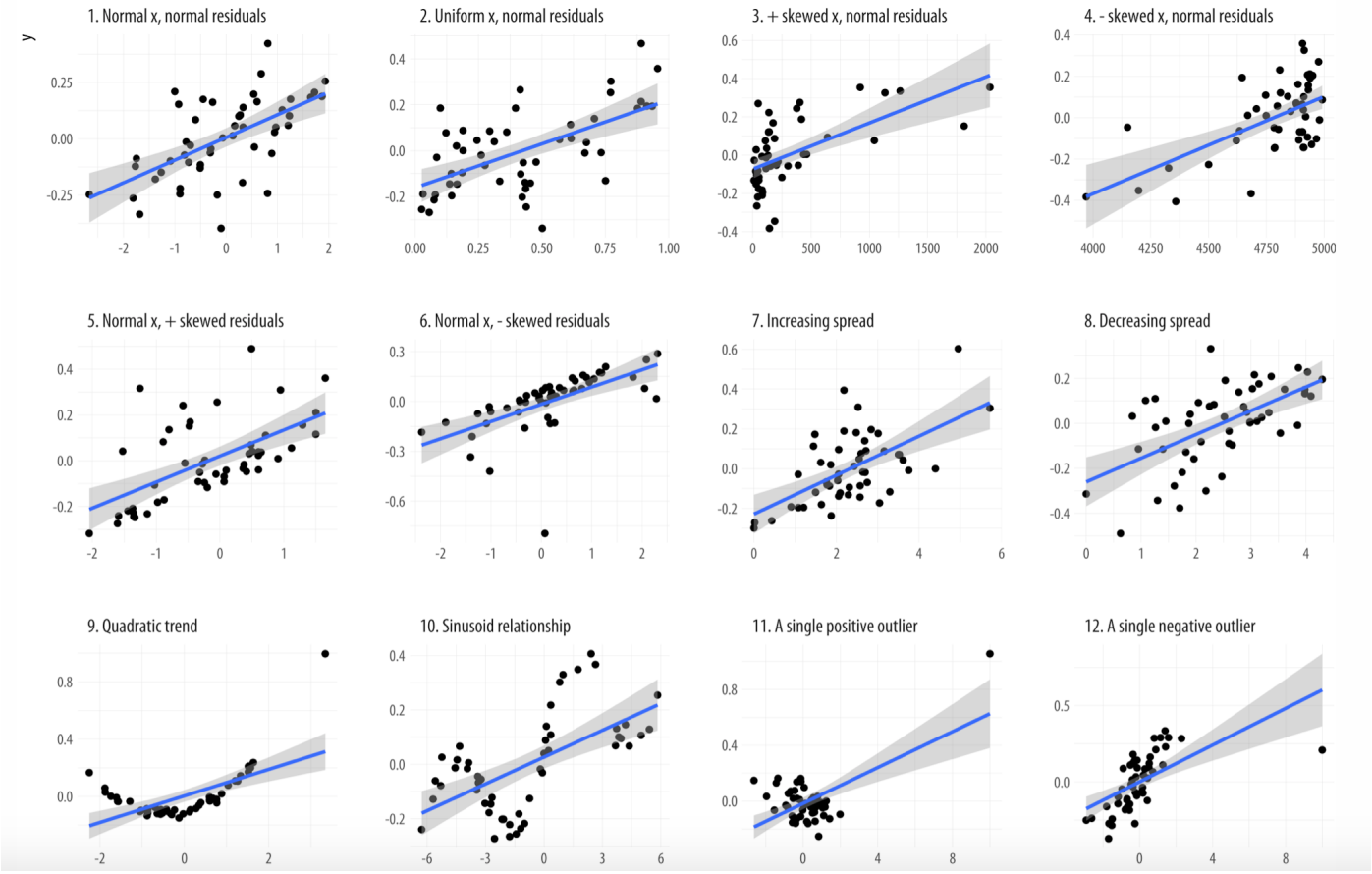
Correlation, denoted by r, is a measure of strength of the *linear* association between two numerical variables.

- r lies between -1 and 1, inclusive
- r equals 1 if and only if all points lie on a line with positive slope
- r equals -1 if and only if all points lie on a line with negative slope



Go to the website `http://guessthecorrelation.com`
You have 7 minutes. Take a screenshot of your highest score, and type your score in the zoom chat. The one with the highest of all will get a prize.

The correlation is only useful if the relationship is linear and there are no outliers.
All of the plots below have a correlation $= 0.6$



The correlation is only useful if the relationship is linear and there are no outliers.

If those conditions are met, it summarizes:

- Strength:
  - Strong relationship if correlation is close to 1 or -1
  - Weak relationship if correlation is close to 0
- Direction:
  - Positive association if correlation is positive
  - Negative association if correlation is negative.

I won't ask you to compute correlation by hands. We can compute correlation in R easily.

Suppose we want to calculate the correlation between 'petal length' and 'petal width' in the following data set with measurements on 150 'iris' flowers.

We can use the following code

```
iris %>%
    select(petal_length, petal_width) %>%
    cor()
```

Output:

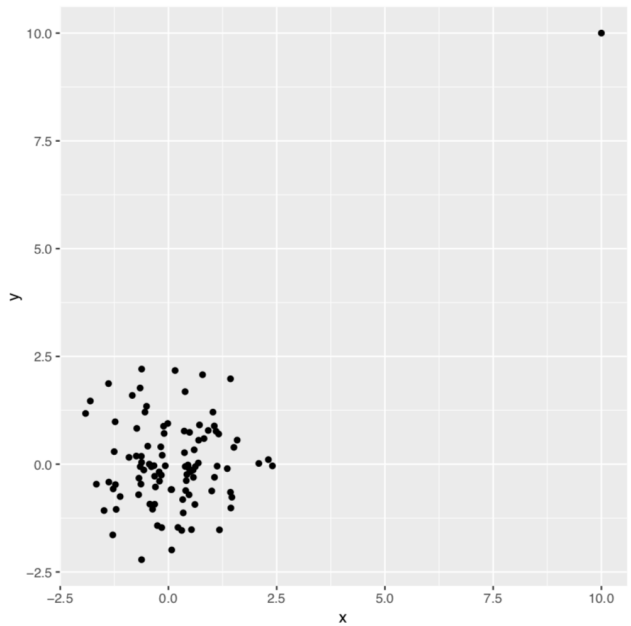
```
##                petal_length petal_width
## petal_length    1.0000000   0.9628654
## petal_width     0.9628654   1.0000000
```

R won't check for you if the relationship is linear and there are no outliers. You must check this yourself.

Poll question

Is the correlation useful here?

a. Yes
b. No

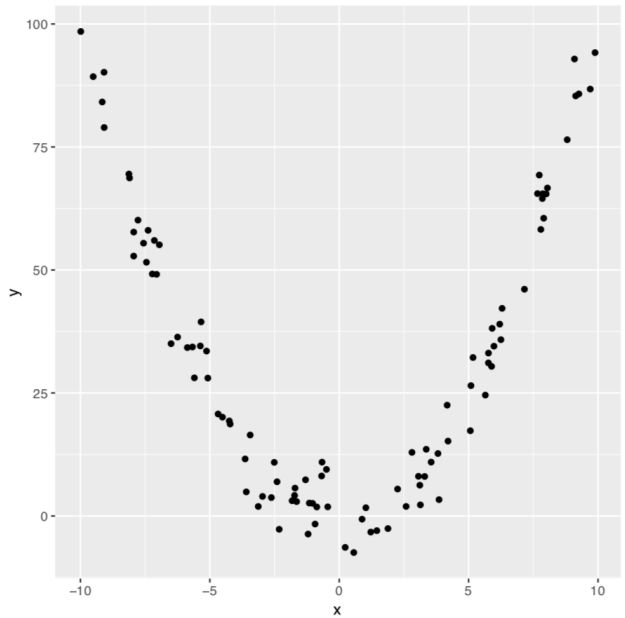


R won't check for you if the relationship is linear and there are no outliers. You must check this yourself.

Poll question

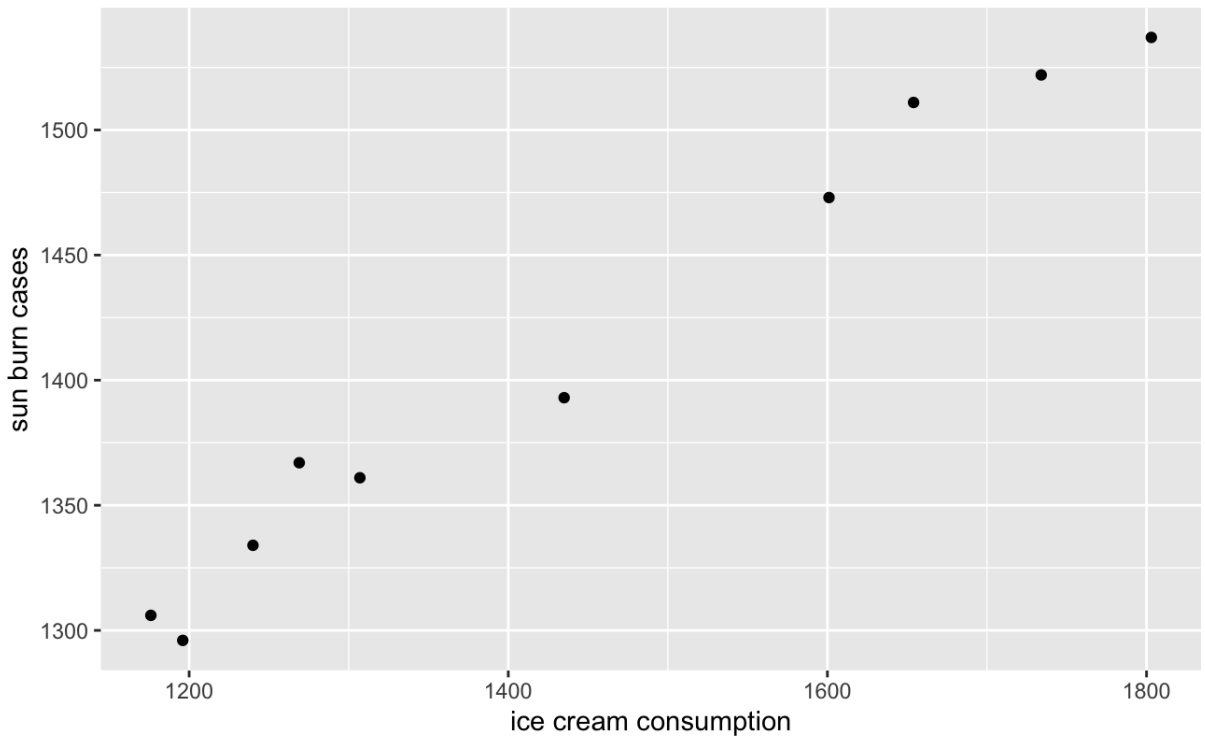
Is the correlation useful here?

a. Yes
b. No



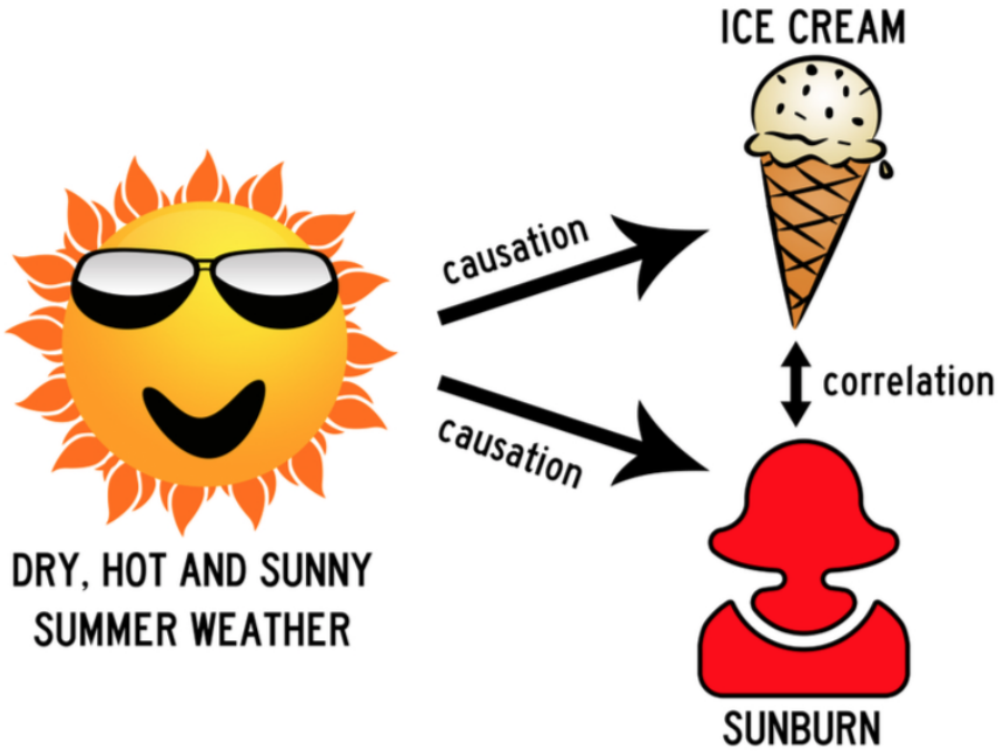
Correlation is not causation!

Just because there is a linear relationship between two variables does not mean we have evidence that one variable causes the other.



Correlation is not causation!

Just because there is a linear relationship between two variables does not mean we have evidence that one variable causes the other.



Even if there really was a cause-and-effect relationship, with correlation we cannot say which variable is the cause and which is the effect. It's also possible that there exists some other unmeasured variable affecting the linear relationship we observe.

And of course, any apparent relationship may be due to nothing more than random chance.

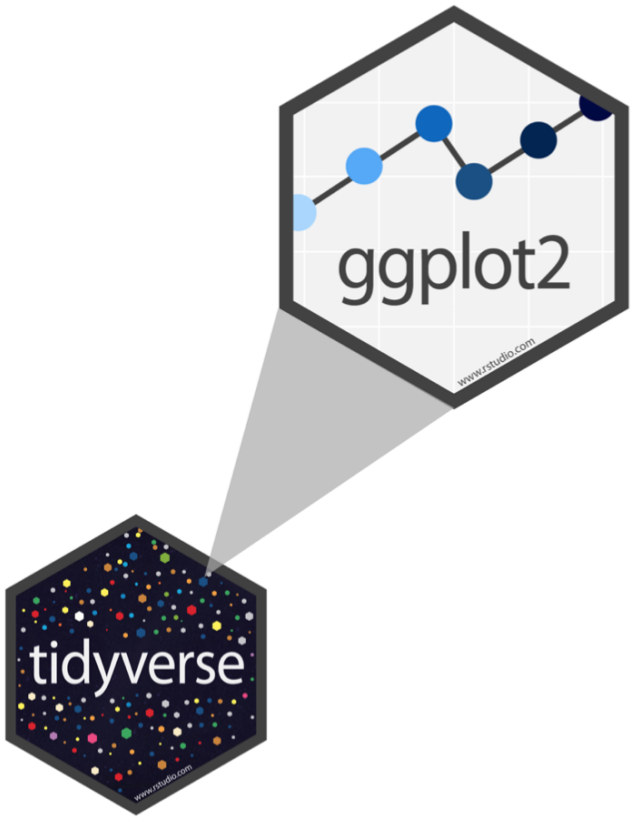
Go to `http://www.tylervigen.com/spurious-correlations`
Browse through some of the plots there. Become fully and deeply convinced that if two variables have a high correlation, that does not tell you anything about one variable causing the other.

Outline

1. Today: Relationship between two numerical variables

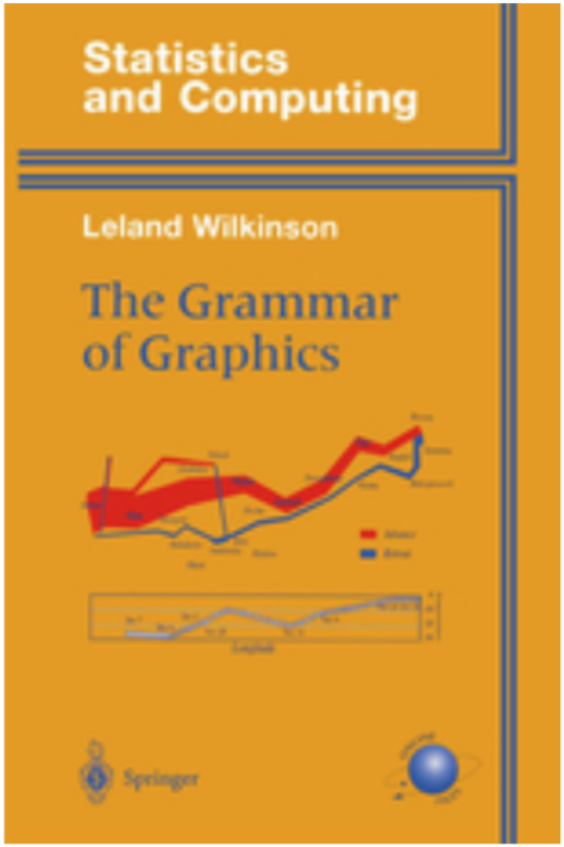
2. Main ideas
    1. Scatterplot
    2. Correlation
    3. Make a scatterplot with ggplot

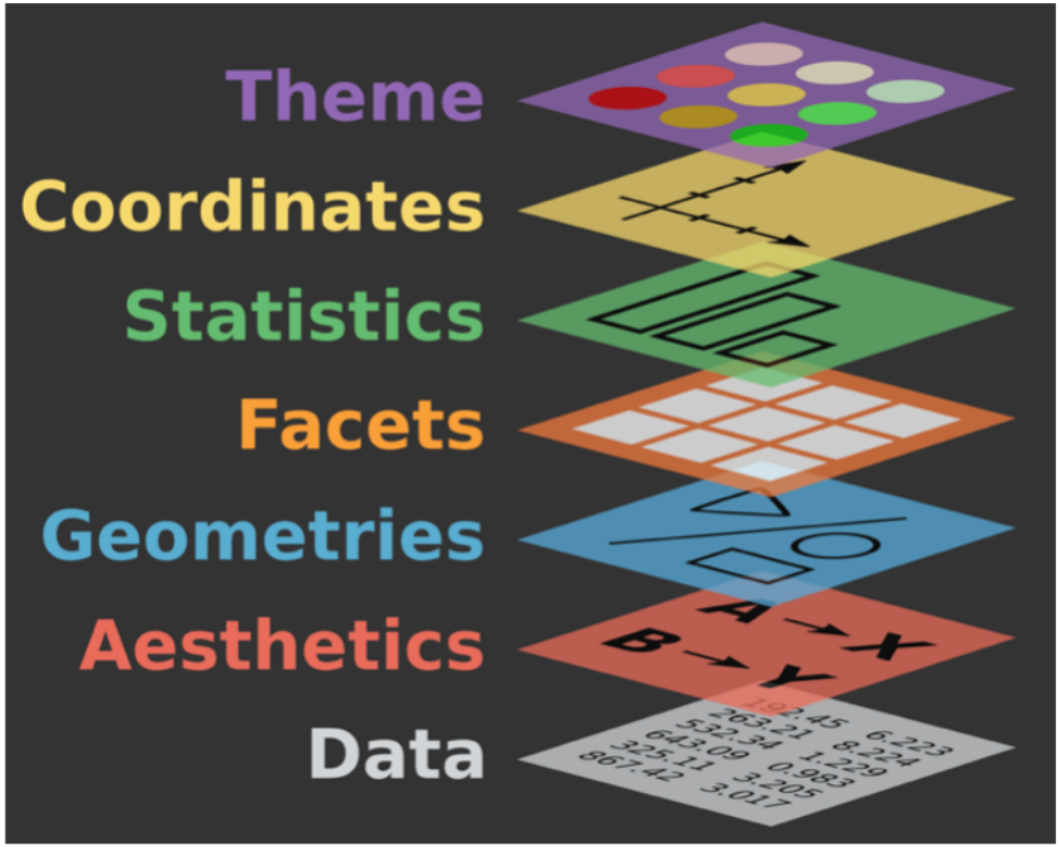
3. Summary



- ggplot2 is tidyverse's data visualization package
- The gg in "ggplot2" stands for Grammar of Graphics
- It is inspired by the book Grammar of Graphics by Leland Wilkinson

A grammar of graphics is a tool that enables us to concisely describe the components of a graphic





ggplot() is the main function in ggplot2

Structure of the code for plots can be summarized as

```
ggplot(data = [d], mapping = aes(x = [x-var], y = [y-var])) +
  geom_xxx()
```

A statistical graphic is a **mapping** of **data** variables to **aes**thetic attributes of **geo**metric objects.

To use ggplot2 functions, first load tidyverse

```
library(tidyverse)
```

```
ggplot(data = [d], mapping = aes(x = [x-var], y = [y-var])) +
    geom_xxx()
```

First, we need to tell R that we want to create a ggplot. This is done by using the **ggplot( )** function.

```
ggplot(data = [d], mapping = aes(x = [x-var], y = [y-var])) +
    geom_xxx()
```

First, we need to tell R that we want to create a ggplot. This is done by using the **ggplot( )** function.

Within the parentheses, we can specify the data frame that contains what we want to plot, using the option **data = [d]**.

```
ggplot(data = [d], mapping = aes(x = [x-var], y = [y-var])) +
    geom_xxx()
```

First, we need to tell R that we want to create a ggplot. This is done by using the **ggplot( )** function.

Within the parentheses, we can specify the data frame that contains what we want to plot, using the option **data = [d]**.

We also have to tell ggplot what columns of the data frame to actually plot – we do this with the argument that stands for aesthetics: **aes( )**.

```
ggplot(data = [d], mapping = aes(x = [x-var], y = [y-var])) +
    geom_xxx()
```

First, we need to tell R that we want to create a ggplot. This is done by using the **ggplot( )** function.

Within the parentheses, we can specify the data frame that contains what we want to plot, using the option **data = [d]**.

We also have to tell ggplot what columns of the data frame to actually plot – we do this with the argument that stands for aesthetics: **aes( )**.

Finally, add a geom layer, which will determine the type of visual representation that will be used for the data. We use **geom_point**

Outline

1. Today: Relationship between two numerical variables

2. Main ideas
    1. Scatterplot
    2. Correlation
    3. Make a scatterplot with ggplot

3. Summary

1. Scatterplot
2. Correlation
3. Make a scatterplot with ggplot

Import message: Correlation does not imply causation!