What is Data?
Part 2: Patterns & Associations

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

August 29, 2016

Prof. Michael Paul

Prof. William Aspray

Overview

This lecture will…

- look at examples of relationships between variables,
- define positive and negative associations,
- and demonstrate how to plot variables and examine their associations in MiniTab.

Most of today will be done with software.

Representation of data: **matrix**

- Each row is an **observation**

Rows

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Columns

- Each column is a **variable**

Cells

- Each cell is a **value**

Representing data in practice

Now let's recreate this matrix in MiniTab

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

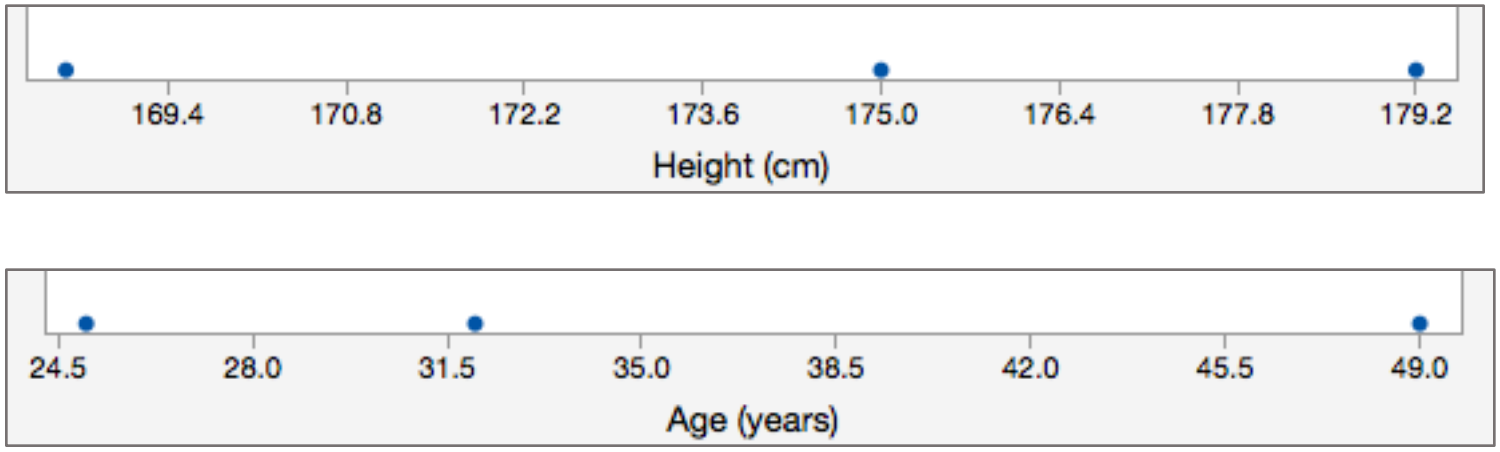
Visualizing data

Dot plots are only for **numerical** variables

Dot plots display the values of **one** variable

Each dot represents an observation

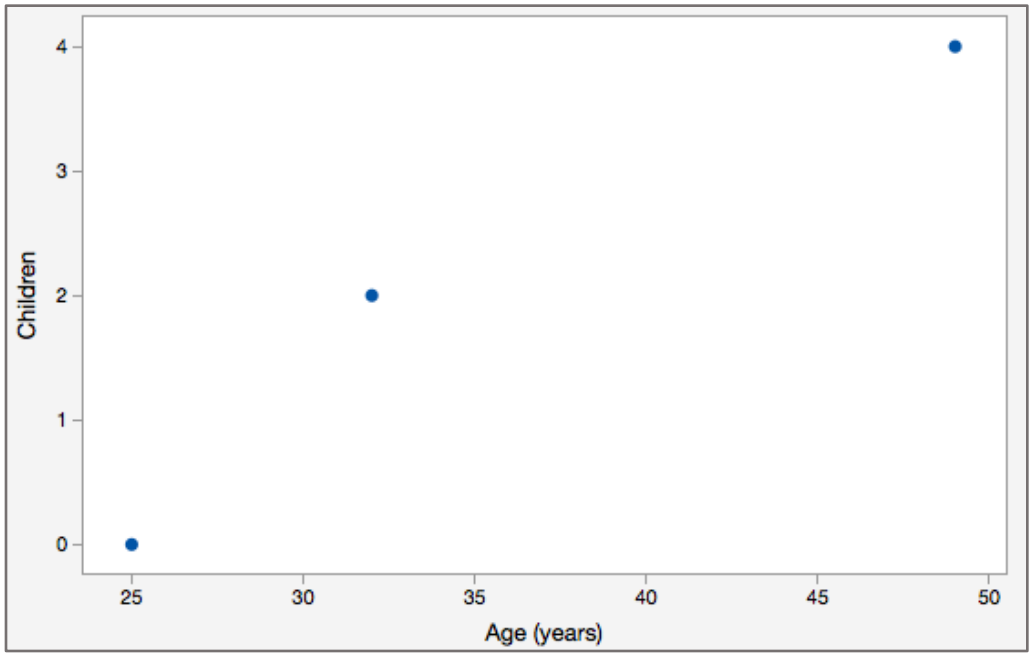
Each dot's position on the x-axis is the value of the variable for that observation



Visualizing data

Scatterplots are an extension of dot plots for **two** variables

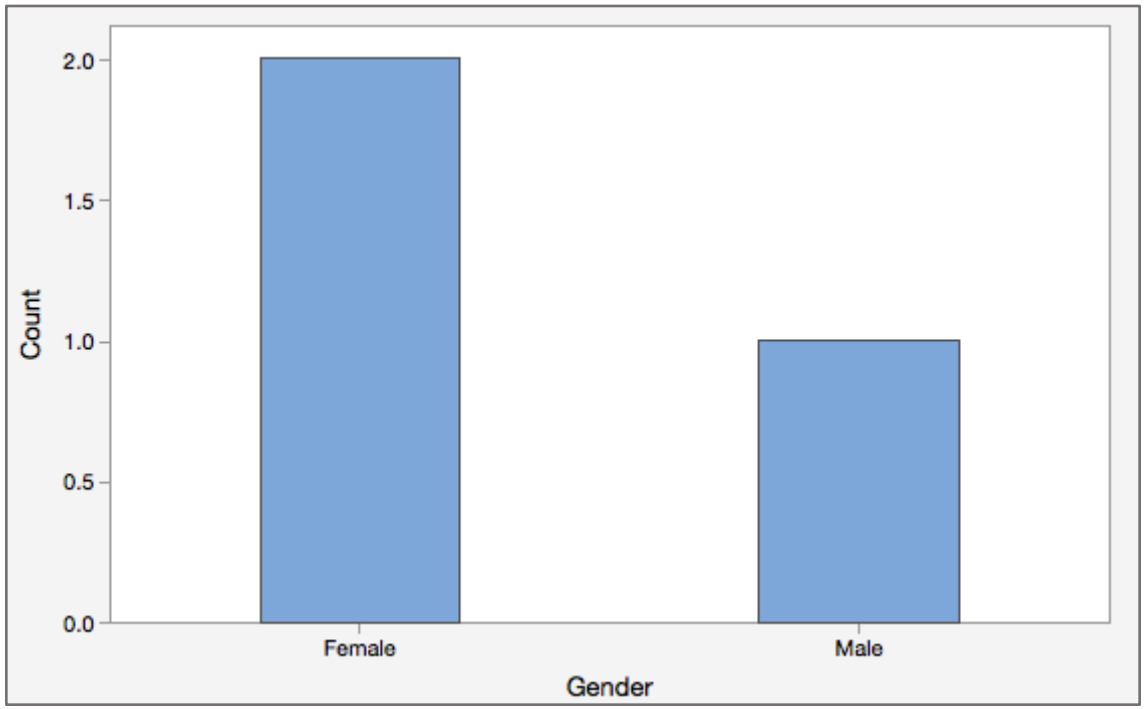
Each dot's position on the x-axis is the value of the first variable; the y-axis is the value of the second



Visualizing data

What about **categorical** data?

We won't get into that today, but there are other options for categories (e.g., **bar charts**)



Relationships between variables

Some variables are related in some way

- Age and number of children
  - The older you are, the more likely you are to have children (in general)

A relationship between variables is called an **association**

Relationships between variables

Example:
Height and weight

- Dataset: measurements of the height and weight of 10,000 children as they grow up
- Association: the taller a child is, the more they will weigh (in general)

Data from:
http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights

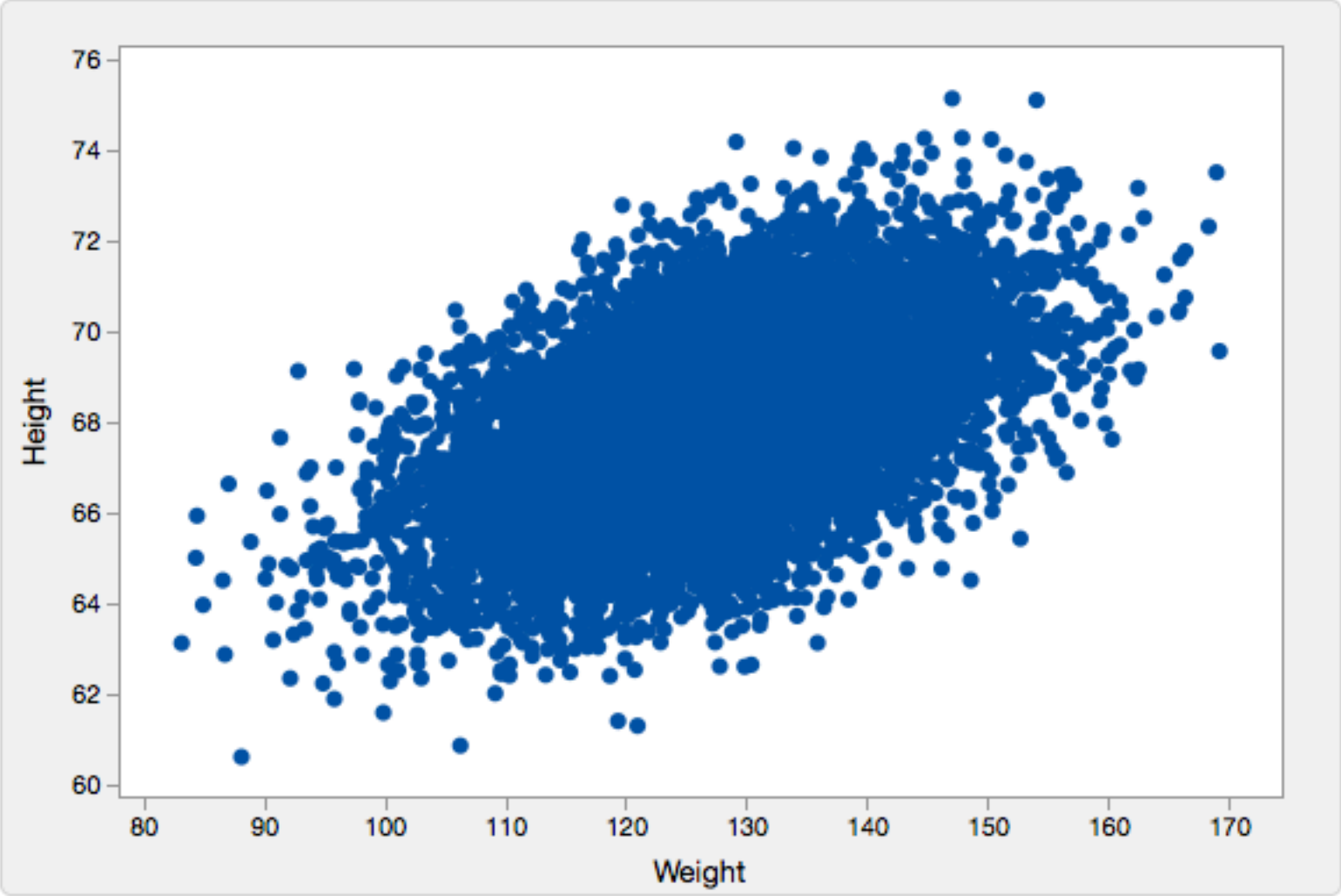
Relationships between variables

Example:
Height and weight

This is called a **positive** association

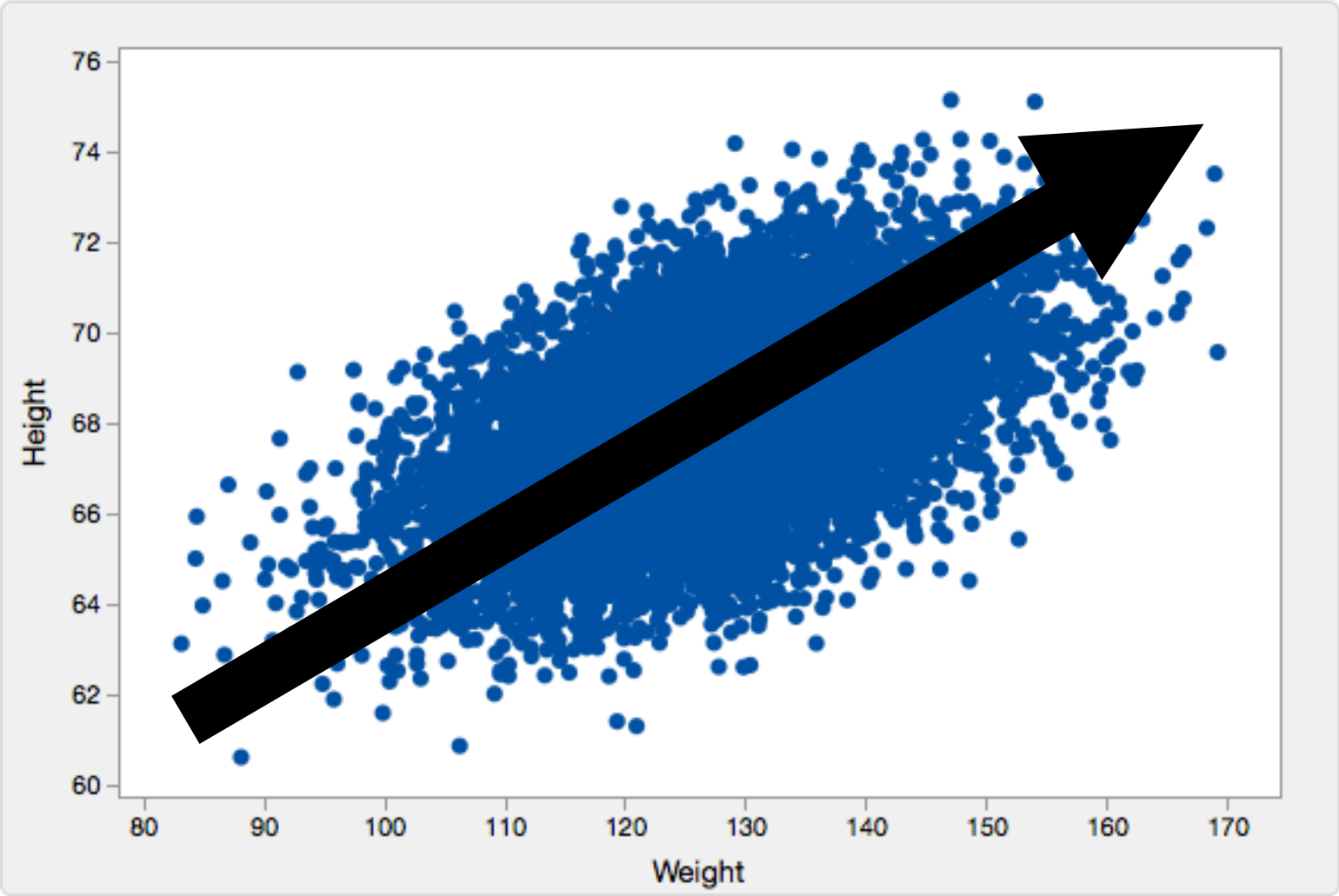
- As the value of one variable increases, the value of the other variable also increases

Visualizing associations



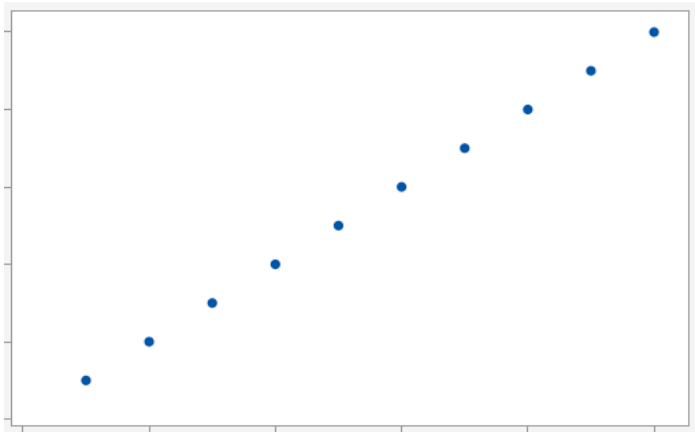
Imagine that the dots in a scatterplot form a line

- If the line is angled upward, the association is positive

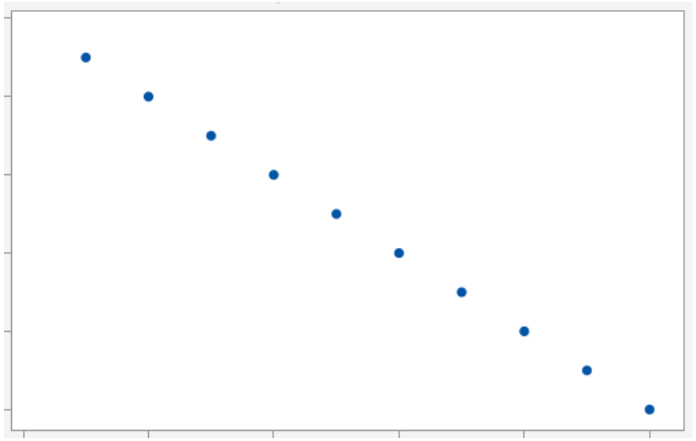


Associations

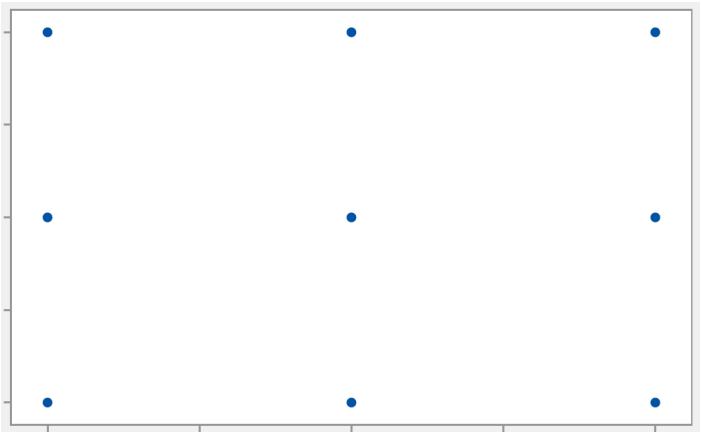
- Positive:



- Negative:



- No association:



Variables that are not associated are called **independent**

Quantifying associations

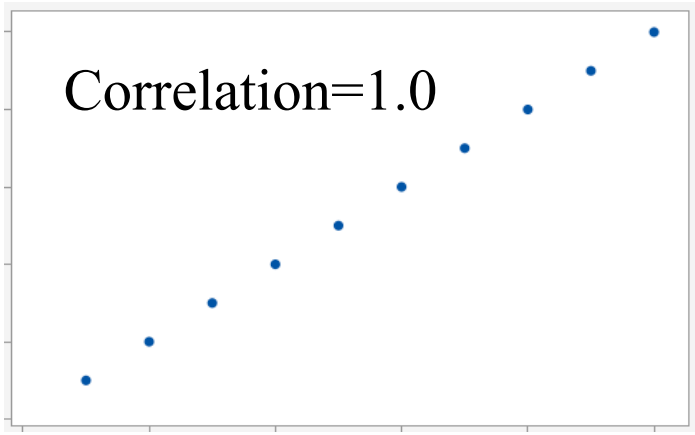
Correlation is a measurement of the association between variables

- Different kinds of correlations
- The correlation we'll use in this class is the **Pearson** correlation
  - When people say "correlation" without specifying, this is what they usually mean
- A correlation is a real number between **[-1, 1]**

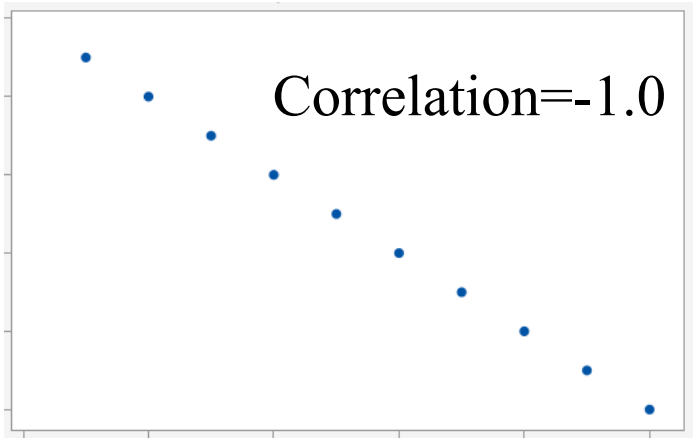
Karl Pearson, 1857-1936

Quantifying associations

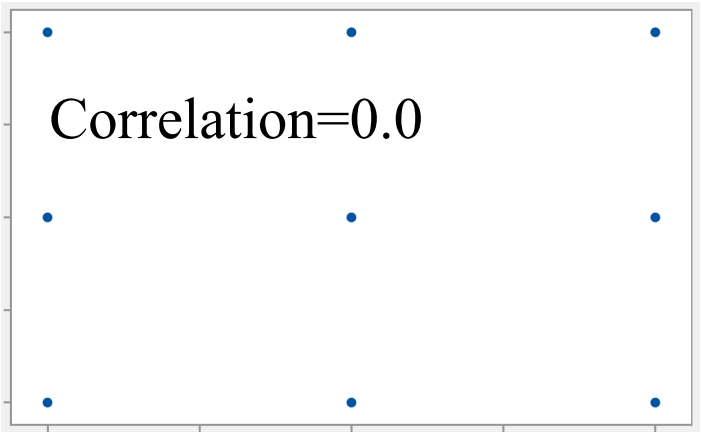
- Positive:



- Negative:



- No association:



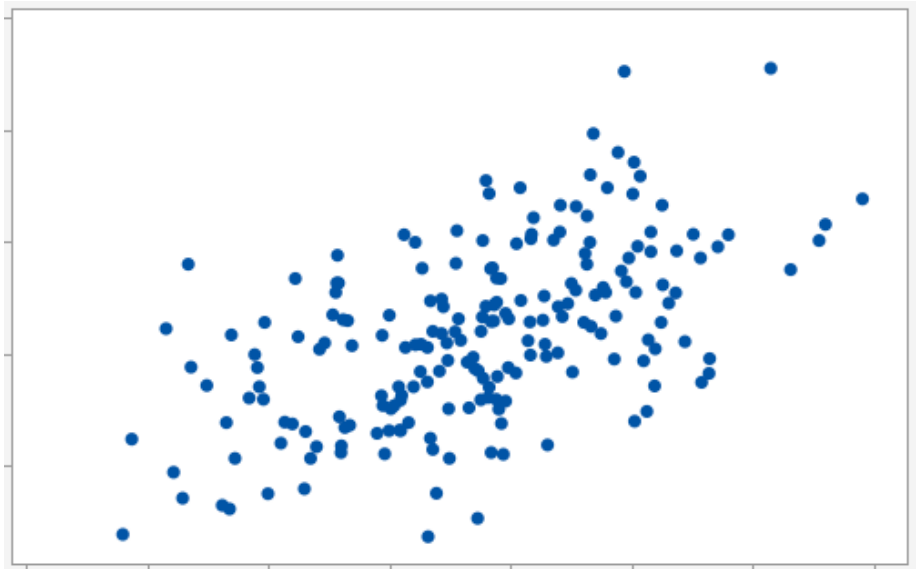
Quantifying associations

Variables that are perfectly associated will have correlations of 1 or -1

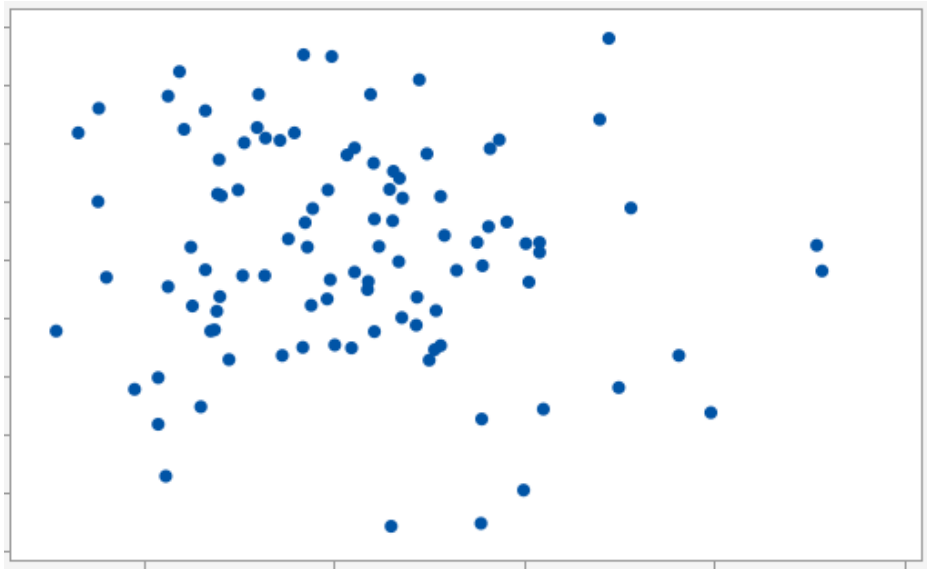
Variables that are independent will have a correlation of 0

In real data, most correlations are somewhere in between

Quantifying associations



Correlation= **0.557**



Correlation= **-0.101**

Your turn

Loan application dataset

http://support.minitab.com/en-us/datasets/basic-statistics-data-sets/loan-applicant-data/

Worksheet column	Description
Income	The income of the applicant, in US dollars
Education	The education level of the applicant, in years
Age	The age of the applicant
Residence	The length of time that the applicant has lived at the current residence, in years
Employ	The length of time that the applicant has worked for the current employer, in years
Savings	The amount of savings that the applicant has, in US dollars
Debt	The amount of debt that the applicant has, in US dollars
Credit cards	The number of credit cards that the applicant has

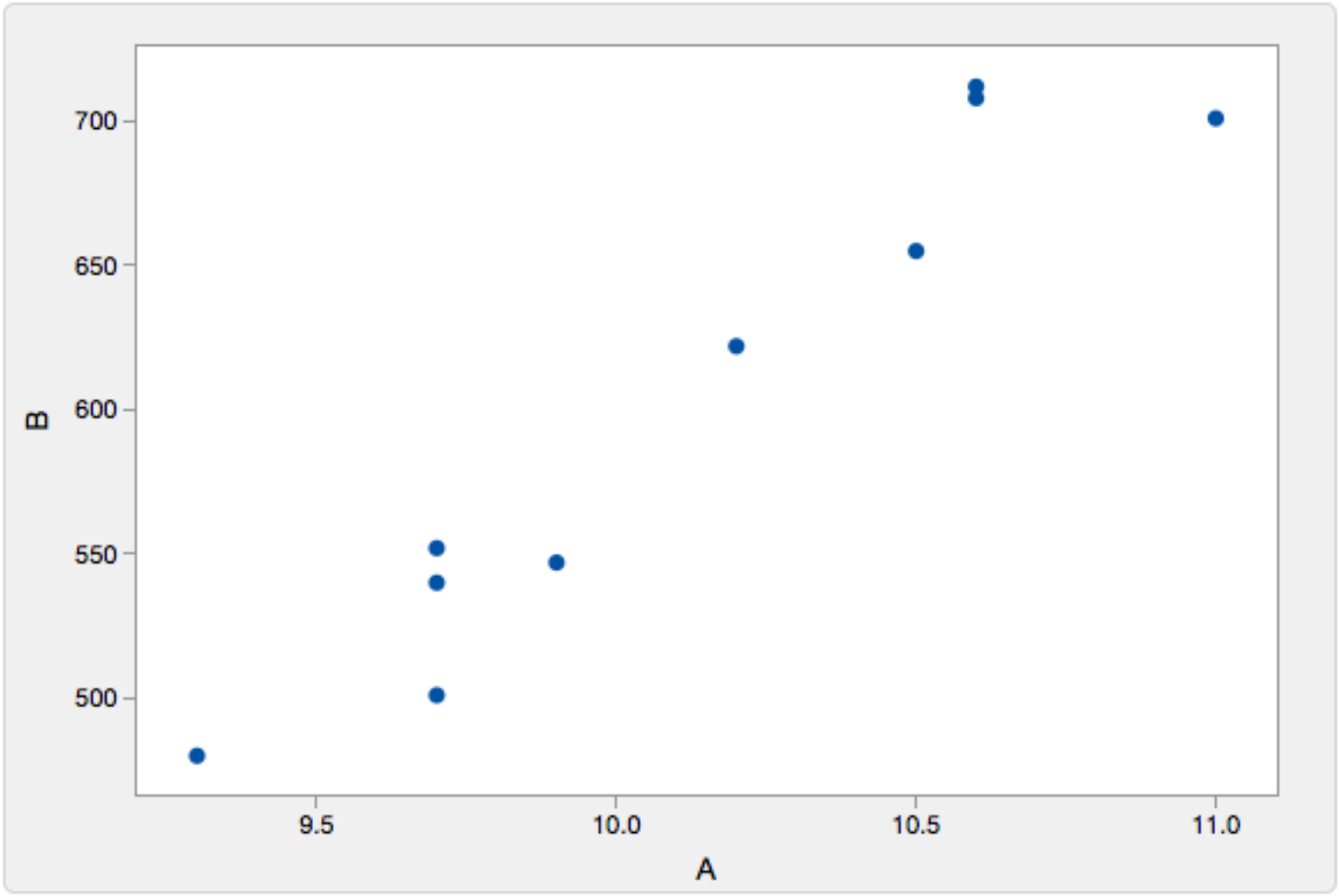
Organize yourselves into groups of 4

Each group should investigate scatterplots and correlations of the following pairs of variables:

- Group 1: Savings, Debt
- Group 2: Employ, Savings
- Group 3: Age, Debt
- Group 4: Education, Credit Cards
- Group 5: Residence, Employ

Each group should also find at least one more pair with an interesting association

One more example



Correlation

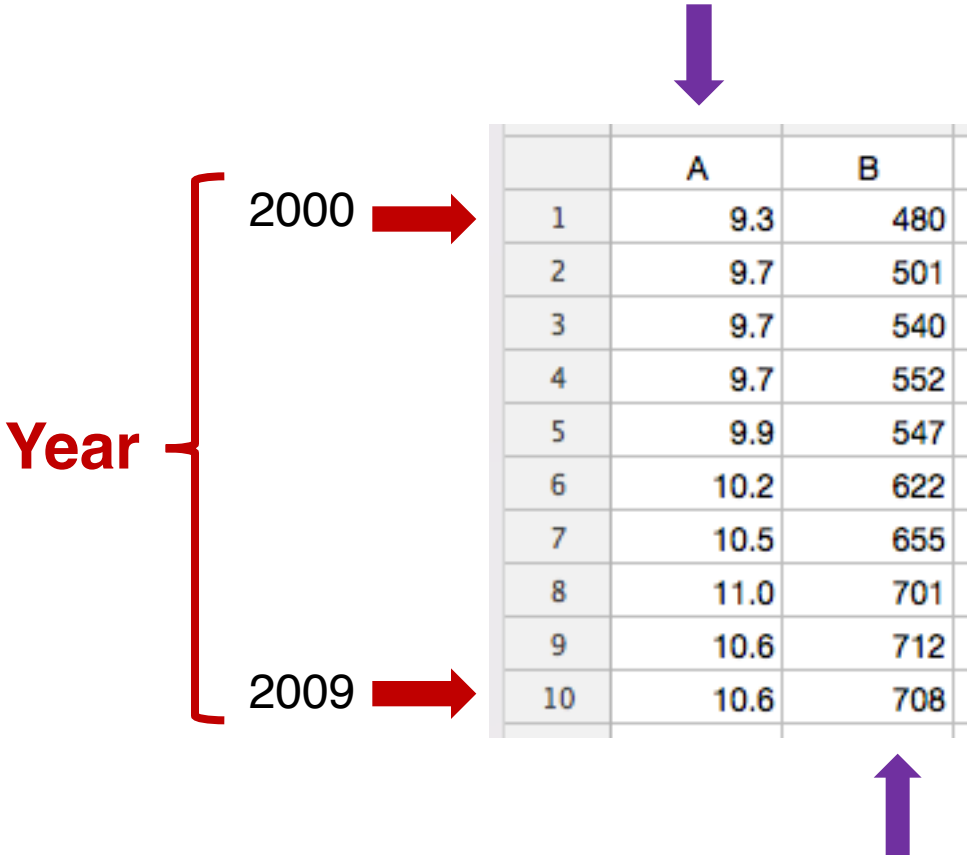
Pearson correlation of A and B = 0.958648
P-Value = <0.0001

How should we interpret this result?

	A	B
1	9.3	480
2	9.7	501
3	9.7	540
4	9.7	552
5	9.9	547
6	10.2	622
7	10.5	655
8	11.0	701
9	10.6	712
10	10.6	708

What do these rows and columns correspond to?

Pounds of mozzarella cheese consumed per capita



	A	B
1	9.3	480
2	9.7	501
3	9.7	540
4	9.7	552
5	9.9	547
6	10.2	622
7	10.5	655
8	11.0	701
9	10.6	712
10	10.6	708

Number of people who earned a PhD in Civil Engineering

Dataset from: http://tylervigen.com/

Spurious correlations

Correlations/associations that are not meaningful – or whose meaning is different than it appears – are said to be **spurious**

"correlation is not causation"

Spurious correlations

Reasons for spurious associations:

- Coincidence
  - Cheese ↔ engineers probably falls into this category
  - Correlations will sometimes happen by chance

Spurious correlations

Reasons for spurious associations:

- Coincidence
- Some other factor in the world is influencing both
  - Researchers have discovered a strong correlation between shark attacks and ice cream sales
  - In this example, summer time explains both variables
    - More people buy ice cream in the summer
    - More people swim in the ocean in the summer
  - In this example, the season (summer) is called a **confounding variable**

Spurious correlations

Reasons for spurious associations:

- Coincidence
- Some other factor in the world is influencing both
- The direction of causation is wrong
  - Sometimes an association is real, but for a different reason than you think
  - Example: healthy people are more likely to have lice than sick people
    - In the Middle Ages, people concluded lice make you healthy
    - Turns out, lice simply don't like to live on sick people

Spurious correlations

Reasons for spurious associations:

- Coincidence
- Some other factor in the world is influencing both
- The direction of causation is wrong

Correlations are interesting and important, but not conclusive

Understanding associations

Why is it useful to measure correlations?

- We can test if associations exist
  - Correlation does not imply causation, but no correlation does imply no causation
- The discovery of associations in big data can lead to new ideas (hypothesis generation)
- Some cases where associations can still inform decisions and predictions
  - People drive faster in red cars – direction of causality doesn't matter to insurance companies