Statistical Methods and Data Analysis I

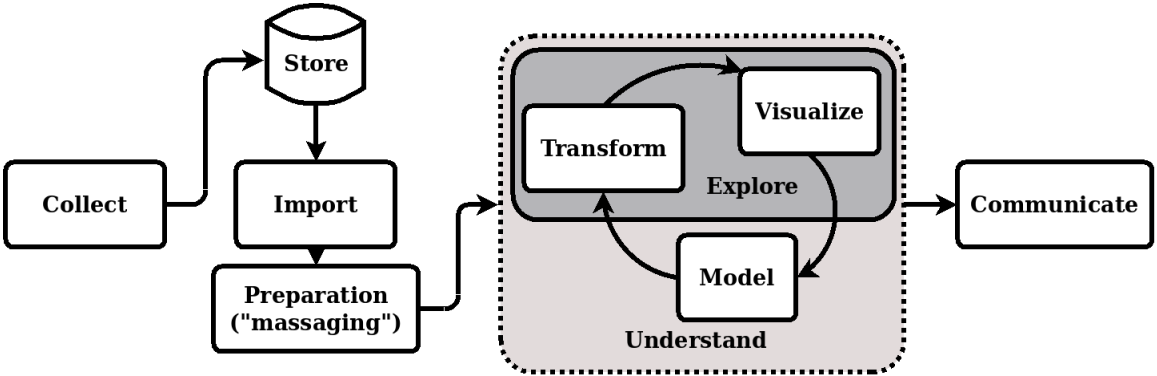
Lecture 2: Data Fundamentals

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

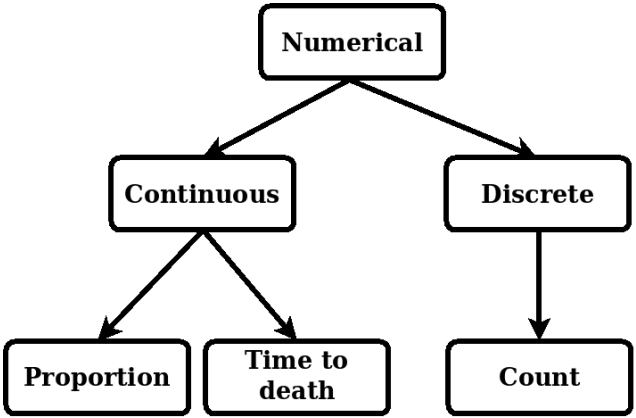
Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

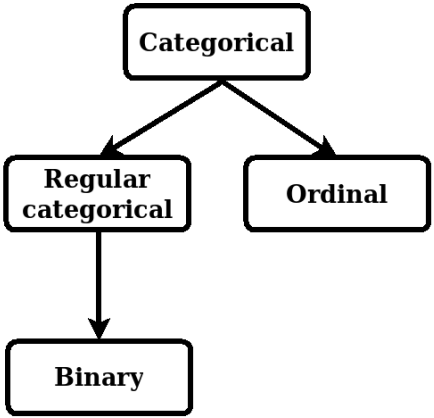
March 05, 2019

Data Lifecycle



Data Types


Numerical
Continuous
Discrete
Proportion
Time to death
Count


Categorical
Regular categorical
Ordinal
Binary

Data Types: Intuition

Data Types: Intuition

- Budget/balance sheet?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular
- Income?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular
- Income? May be numerical or categorical (probably ordinal)

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular
- Income? May be numerical or categorical (probably ordinal)
- Religious affiliation?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular
- Income? May be numerical or categorical (probably ordinal)
- Religious affiliation? Categorical, regular

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular
- Income? May be numerical or categorical (probably ordinal)
- Religious affiliation? Categorical, regular
- Political party affiliation?

Data Types: Intuition

- Budget/balance sheet? Numerical (continuous or discreet?)
- Factory production numbers? Numerical (discreet or continuous)
- Political polls? Numerical (percentages — continuous/proportion, sample counts — discrete/count)
- Age group (13–18, 20–24, 25–35, 35–50, 50–65, 65+)? Categorical, ordinal
- Gender (male/female)? Categorical, binary
- Education (primary, secondary, college, advanced degree)? Categorical, ordinal
- Telephone area code? Categorical, regular
- Geographical (country, region, city)? Categorical, regular
- Income? May be numerical or categorical (probably ordinal)
- Religious affiliation? Categorical, regular
- Political party affiliation? Categorical, regular

Data Types: Example

The `ggplot2::mpg` data set (file `loadmpg.R`)

```
library(tidyverse)
mpg %>% print(n=Inf,width=Inf) # use options(width=120) if needed
```

Data Types: Example

The `ggplot2::mpg` data set (file `loadmpg.R`)

```
library(tidyverse)
mpg %>% print(n=Inf,width=Inf) # use options(width=120) if needed
```

↓ variable

```
    manufacturer model                displ  year   cyl trans      drv     cty   hwy fl    class
    <chr>        <chr>                <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
...
  7 audi         a4                     3.1  2008     6 auto(av)   f        18    27 p     compact
  8 audi         a4 quattro             1.8  1999     4 manual(m5) 4        18    26 p     compact
...
 23 chevrolet    c1500 suburban 2wd     6    2008     8 auto(l4)   r        12    17 r     suv
 24 chevrolet    corvette               5.7  1999     8 manual(m6) r        16    26 p     2seater
...
150 nissan       maxima                 3.5  2008     6 auto(av)   f        19    25 p     midsize
151 nissan       pathfinder 4wd         3.3  1999     6 auto(l4)   4        14    17 r     suv
...
173 subaru       impreza awd            2.5  2008     4 manual(m5) 4        20    27 r     compact
...
198 toyota       corolla                1.8  2008     4 auto(l4)   f        26    35 r     compact
199 toyota       land cruiser wagon 4wd 4.7  1999     8 auto(l4)   4        11    15 r     suv
...
221 volkswagen   jetta                  2.8  1999     6 manual(m5) f        17    24 r     compact
222 volkswagen   new beetle             1.9  1999     4 manual(m5) f        35    44 d     subcompact
...
234 volkswagen   passat                 3.6  2008     6 auto(s6)   f        17    26 p     midsize
```

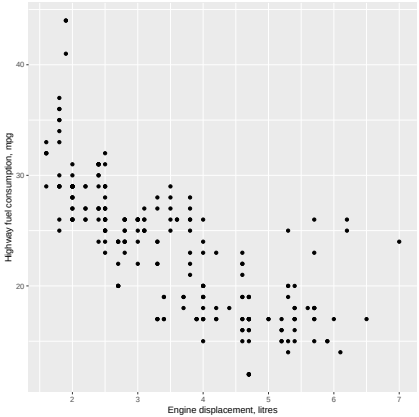
← observation

Independent and Associated variables

- When two variables show some connection with one another, they are called *associated* or *dependent* variables.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.
- There are techniques to determine association (or independence, as the case may be) between variables and we will study them.

Relationship Between Variables

Is highway fuel consumption associated with engine displacement?



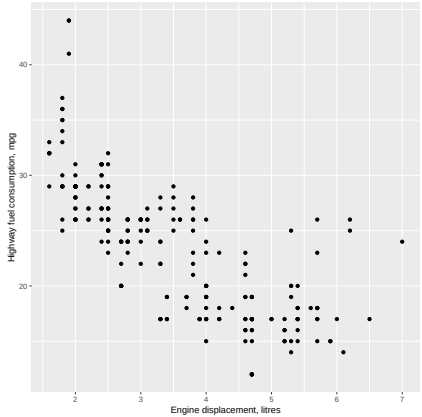
How to get this graph

```
library(tidyverse)

p <- ggplot(data = mpg) +
    geom_point(mapping = aes(x=displ,y=hwy)) +
    xlab('Engine displacement, litres') +
    ylab('Highway fuel consumption, mpg')

ggsave(plot=p,filename='fig/mpg_assoc_displ_hwy_1.pdf',device='pdf')

p <- p + geom_smooth(mapping = aes(x=displ,y=hwy))

ggsave(plot=p,filename='fig/mpg_assoc_displ_hwy_2.pdf',device='pdf')
```

Relationship Between Variables

Is highway fuel consumption associated with engine displacement?

Seems that cars with larger engines use more fuel — negative association...



How to get this graph

```
library(tidyverse)

p <- ggplot(data = mpg) +
    geom_point(mapping = aes(x=displ,y=hwy)) +
    xlab('Engine displacement, litres') +
    ylab('Highway fuel consumption, mpg')

ggsave(plot=p,filename='fig/mpg_assoc_displ_hwy_1.pdf',device='pdf')

p <- p + geom_smooth(mapping = aes(x=displ,y=hwy))

ggsave(plot=p,filename='fig/mpg_assoc_displ_hwy_2.pdf',device='pdf')
```

Relationship Between Variables

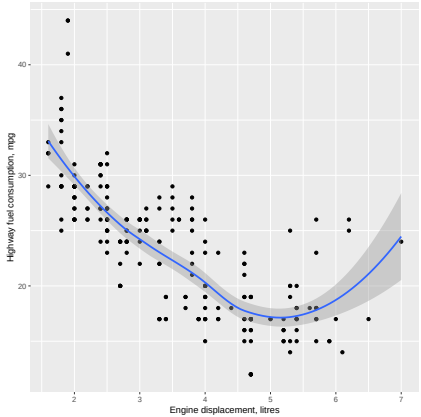
Is highway fuel consumption associated with engine displacement?

Seems that cars with larger engines use more fuel — negative association...

Seen even better with a bit of smoothing...

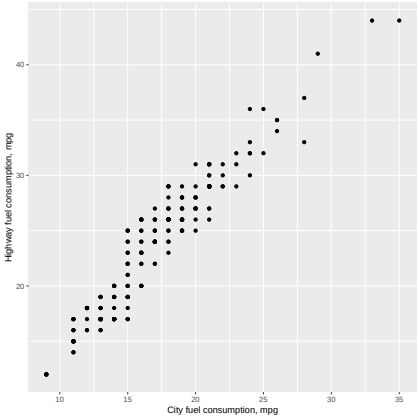
How to get this graph

```
library(tidyverse)

p <- ggplot(data = mpg) +
    geom_point(mapping = aes(x=displ,y=hwy)) +
    xlab('Engine displacement, litres') +
    ylab('Highway fuel consumption, mpg')

ggsave(plot=p,filename='fig/mpg_assoc_displ_hwy_1.pdf',device='pdf')

p <- p + geom_smooth(mapping = aes(x=displ,y=hwy))

ggsave(plot=p,filename='fig/mpg_assoc_displ_hwy_2.pdf',device='pdf')
```



Relationship Between Variables

Is highway fuel consumption associated with city performance?



How to get this graph

```
library(tidyverse)

p <- ggplot(data = mpg) +
    geom_point(mapping = aes(x=cty,y=hwy)) +
    xlab('City fuel consumption, mpg') +
    ylab('Highway fuel consumption, mpg')

ggsave(plot=p,filename='fig/mpg_assoc_cty_hwy_1.pdf',device='pdf')

p <- p + geom_smooth(mapping = aes(x=cty,y=hwy))

ggsave(plot=p,filename='fig/mpg_assoc_cty_hwy_2.pdf',device='pdf')
```

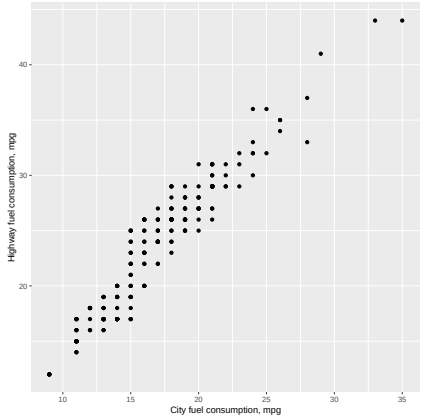
Relationship Between Variables

Is highway fuel consumption associated with city performance?

It rather stands to reason that these two are positively associated...

How to get this graph

```
library(tidyverse)

p <- ggplot(data = mpg) +
    geom_point(mapping = aes(x=cty,y=hwy)) +
    xlab('City fuel consumption, mpg') +
    ylab('Highway fuel consumption, mpg')

ggsave(plot=p,filename='fig/mpg_assoc_cty_hwy_1.pdf',device='pdf')

p <- p + geom_smooth(mapping = aes(x=cty,y=hwy))

ggsave(plot=p,filename='fig/mpg_assoc_cty_hwy_2.pdf',device='pdf')
```



Relationship Between Variables

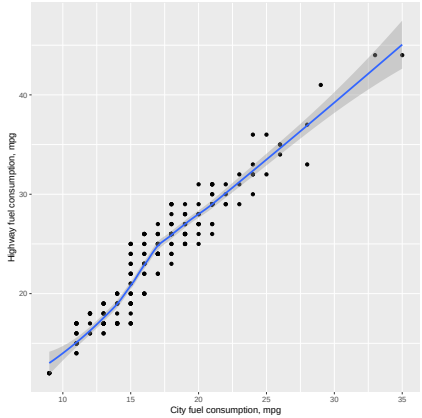
Is highway fuel consumption associated with city performance?

It rather stands to reason that these two are positively associated...

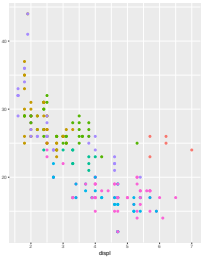
Again, seen even better with a bit of smoothing...

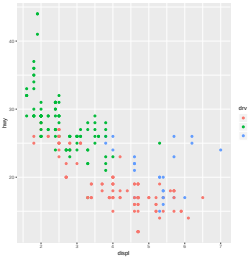
How to get this graph

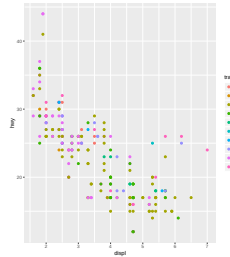
```
library(tidyverse)

p <- ggplot(data = mpg) +
    geom_point(mapping = aes(x=cty,y=hwy)) +
    xlab('City fuel consumption, mpg') +
    ylab('Highway fuel consumption, mpg')

ggsave(plot=p,filename='fig/mpg_assoc_cty_hwy_1.pdf',device='pdf')

p <- p + geom_smooth(mapping = aes(x=cty,y=hwy))

ggsave(plot=p,filename='fig/mpg_assoc_cty_hwy_2.pdf',device='pdf')
```

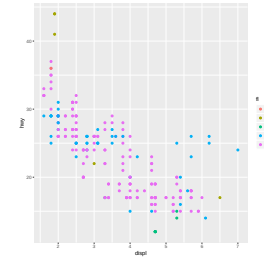


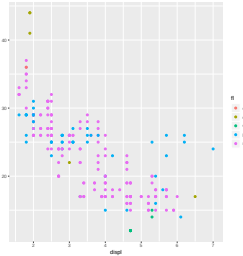
Exploring The mpg Data Set

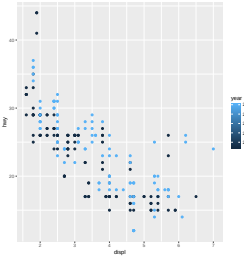


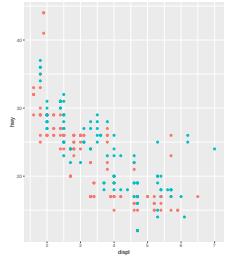


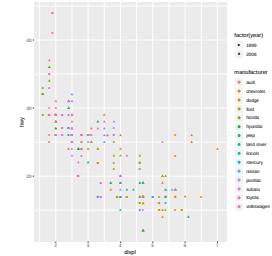












Toolkit: How To Get The Graphs On The Previous slide

Exploring `ggplot2::mpg` (file `explore_mpg.R`)

```
library(tidyverse)

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=class))
ggsave(plot=p,filename='fig/explore_mpg_class.pdf',device='pdf')

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=drv))
ggsave(plot=p,filename='fig/explore_mpg_drv.pdf',device='pdf')

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=trans))
ggsave(plot=p,filename='fig/explore_mpg_trans.pdf',device='pdf')

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=manufacturer))
ggsave(plot=p,filename='fig/explore_mpg_manufacturer.pdf',device='pdf')

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=fl))
ggsave(plot=p,filename='fig/explore_mpg_fl.pdf',device='pdf')

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=year))
ggsave(plot=p,filename='fig/explore_mpg_year.pdf',device='pdf')

p <- ggplot(data=mpg) + geom_point(mapping=aes(x=displ,y=hwy,color=factor(year)))
ggsave(plot=p,filename='fig/explore_mpg_fyear.pdf',device='pdf')

p <- ggplot(data=mpg) +
    geom_point(mapping=aes(x=displ,y=hwy,shape=factor(year),color=manufacturer))
ggsave(plot=p,filename='fig/explore_mpg_maker_year.pdf',device='pdf')
```

Explanatory and Response Variables

Given 2 variables, identify which of the two is **suspected** of affecting the other:

$$\text{explanatory variable} \xrightarrow{\textit{might affect}} \text{response variable}$$

Labeling variables as explanatory and response does not indicate a causal relationship between them, even if there is an observed association between the two variables.

(NB: we will discuss correlation and causation later on.)

Such labeling reflects an assumption, often implicit, and certainly not a known fact. It may be regarded a **hypothesis** or may be considered a part of a **model**.

Observations and Experiments

- Observational studies
  - Researchers collect data in ways that do not interfere with how the data are generated. In other words, they merely “observe”.
  - Observational studies can only establish associations between explanatory and observed variables.
- Experimental studies
  - Researchers assign subjects to various “treatments” and study the “effects”.
  - “Treatments” are designed to manipulate explanatory variables, and the “effects” correspond to changes in the response variables.
  - Various techniques related to such assignment (randomization, blind and double-blind studies, etc.) will be studied later.
  - So will techniques of performing “controlled experiments” (changing only one explanatory variable at a time, etc.).
  - The intent is to attempt to establish a causal relationship (or lack thereof) between explanatory and response variables.
  - This may or may not succeed. Various pitfalls will also be discussed in this course.

Prospective and Retrospective Studies

- A *prospective* study identifies subjects and collects information on an ongoing basis.
  - Example: the Nurses' Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
- A *retrospective* study collects data *post factum*, after relevant events have taken place.
  - Example: researchers may review past medical records.

The Most Important Meal of the Day

New study sponsored by General Mills says that eating breakfast makes girls thinner

Study: Breakfast Helps Girls Stay Slim
I love these studies....and finding out who sponsored them!

By ALEX DOMINGUEZ, Associated Press

Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills.

"Not eating breakfast is the worst thing you can do, that's really the take-home message for teenage girls," said study author Bruce Barton, the Maryland institute's president and CEO.

The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

“Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Prospective or retrospective?

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Prospective or retrospective?

Prospective...

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Prospective or retrospective?

Prospective...

What is the conclusion?

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Prospective or retrospective?

Prospective...

What is the conclusion?

There is an **association** *between girls eating breakfast and being slimmer.*

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Prospective or retrospective?

Prospective...

What is the conclusion?

There is an **association** *between girls eating breakfast and being slimmer.*

Bonus question: is any disclosure necessary?

Breakfast Study: Analysis

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an **observational study** *since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.*

Prospective or retrospective?

Prospective...

What is the conclusion?

There is an **association** *between girls eating breakfast and being slimmer.*

Bonus question: is any disclosure necessary?

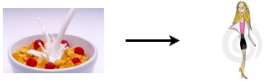
The study was sponsored by General Mills, a Minnesota-based manufacturer of branded foods, including Yoplait, Pillsbury, Häagen-Dazs, Cheerios, Trix, Cocoa Puffs, and Lucky Charms.

Cereals are good for you, eh?

Possible Explanations

Possible Explanations

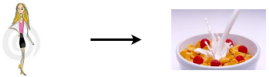
1. Eating breakfast causes girls to be thinner.



Possible Explanations

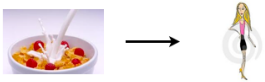
1. Eating breakfast causes girls to be thinner.

2. Being thin causes girls to eat breakfast.

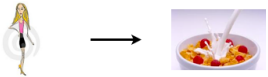


Possible Explanations

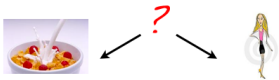
1. Eating breakfast causes girls to be thinner.



2. Being thin causes girls to eat breakfast.



3. There may be a common cause — a *confounding* variable that affects both the explanatory and the response variable and that makes it seem like there is a relationship between the two.



Source: https://github.com/OpenIntroOrg/openintro-statistics-slides.git

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.
- More people attended Obama's inauguration ceremony than Trump's. However, **many** more people watched Trump's inauguration on streaming video than Obama's.
  What does it say about Trump's popularity?

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.
- More people attended Obama's inauguration ceremony than Trump's. However, **many** more people watched Trump's inauguration on streaming video than Obama's.
  What does it say about Trump's popularity?
  Streaming video was not so commonplace at the time of Obama's inauguration.

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.
- More people attended Obama's inauguration ceremony than Trump's. However, **many** more people watched Trump's inauguration on streaming video than Obama's.
  What does it say about Trump's popularity?
  Streaming video was not so commonplace at the time of Obama's inauguration.
  Maybe, given the availability of streaming video, lots of people decided not to attend Trump's inauguration in person.

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.
- More people attended Obama's inauguration ceremony than Trump's. However, **many** more people watched Trump's inauguration on streaming video than Obama's.
  What does it say about Trump's popularity?
  Streaming video was not so commonplace at the time of Obama's inauguration.
  Maybe, given the availability of streaming video, lots of people decided not to attend Trump's inauguration in person.
  Overall, hard to tell without a more detailed analysis.

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.
- More people attended Obama's inauguration ceremony than Trump's. However, **many** more people watched Trump's inauguration on streaming video than Obama's.
  What does it say about Trump's popularity?
  Streaming video was not so commonplace at the time of Obama's inauguration.
  Maybe, given the availability of streaming video, lots of people decided not to attend Trump's inauguration in person.
  Overall, hard to tell without a more detailed analysis.
- At least before the advent of mobile phones there was a very high correlation between the load on telephone lines in Washington DC and the level of water in the Potomac river.
  What the...?

Examples of Confounding Variables

- There is a high correlation between the birth rate and stork populations across European countries.
  Do storks bring babies?
  In larger countries there are more storks and more babies are born.
- More people attended Obama's inauguration ceremony than Trump's. However, **many** more people watched Trump's inauguration on streaming video than Obama's.
  What does it say about Trump's popularity?
  Streaming video was not so commonplace at the time of Obama's inauguration.
  Maybe, given the availability of streaming video, lots of people decided not to attend Trump's inauguration in person.
  Overall, hard to tell without a more detailed analysis.
- At least before the advent of mobile phones there was a very high correlation between the load on telephone lines in Washington DC and the level of water in the Potomac river.
  What the...?
  People call each other when it rains.

Fallacy: *Post hoc propter ergo hoc*

- Storks don't bring babies. *Do* **infant seats** *bring babies?*

Fallacy: *Post hoc propter ergo hoc*

- Storks don't bring babies. *Do* **infant seats** *bring babies?*
  People shop for infant seats before babies are born.

Fallacy: *Post hoc propter ergo hoc*

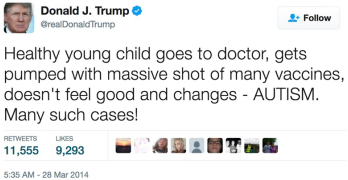
- Storks don't bring babies. *Do* **infant seats** *bring babies?*
  People shop for infant seats before babies are born.
  Pregnancy confounds things.

Fallacy: *Post hoc propter ergo hoc*

- Storks don't bring babies. *Do* **infant seats** *bring babies?*
  People shop for infant seats before babies are born.
  Pregnancy confounds things.
- *Does vaccination cause autism?*

Fallacy: *Post hoc propter ergo hoc*

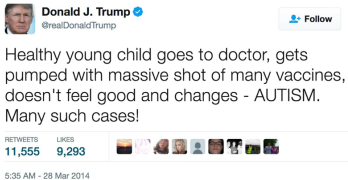
- Storks don't bring babies. *Do* **infant seats** *bring babies? People shop for infant seats before babies are born. Pregnancy confounds things.*
- *Does vaccination cause autism? Many people think so. Ongoing research trying to refute the notion convincingly.*



Credit: https://callingbullshit.org

Fallacy: *Post hoc propter ergo hoc*

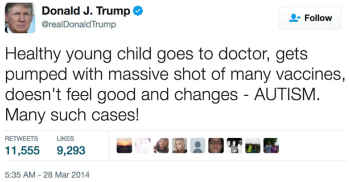
- Storks don't bring babies. *Do* **infant seats** *bring babies? People shop for infant seats before babies are born. Pregnancy confounds things.*
- *Does vaccination cause autism? Many people think so. Ongoing research trying to refute the notion convincingly.*
- *Do speed cameras lead to a decrease in number of traffic accidents?*



Credit: https://callingbullshit.org

Fallacy: *Post hoc propter ergo hoc*

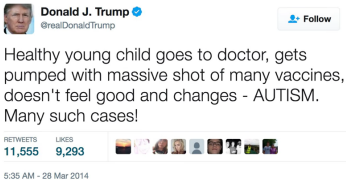
- Storks don't bring babies. *Do* **infant seats** *bring babies? People shop for infant seats before babies are born. Pregnancy confounds things.*
- *Does vaccination cause autism? Many people think so. Ongoing research trying to refute the notion convincingly.*
- *Do speed cameras lead to a decrease in number of traffic accidents? Infrastructure improvements and newer cars do.*



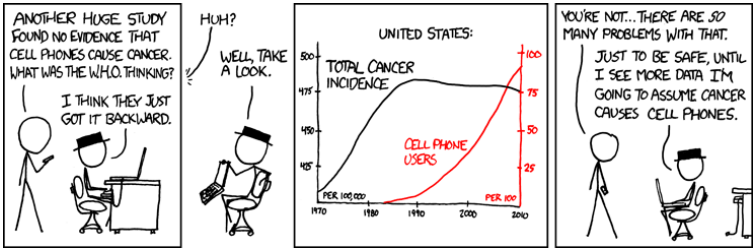
Credit: https://callingbullshit.org

Fallacy: *Post hoc propter ergo hoc*

- Storks don't bring babies. *Do* **infant seats** *bring babies? People shop for infant seats before babies are born. Pregnancy confounds things.*
- *Does vaccination cause autism? Many people think so. Ongoing research trying to refute the notion convincingly.*
- *Do speed cameras lead to a decrease in number of traffic accidents? Infrastructure improvements and newer cars do.*



Credit: https://callingbullshit.org



Credit: https://xkcd.com/925/