GENERAL ASSEMBLY

UNDERSTANDING YOUR DATA:

CORRELATION

Jim Byers, Business Intelligence Manager

UNDERSTANDING YOUR DATA: CORRELATION

Learning Objectives

At the end of this module you will be able to:

- Describe what correlation is and provide an example of positive and negative correlation
- Be able to complete this phrase “Correlation does not imply ____!”
- Use Pandas to look at the data, create a plot of the data and determine the correlation coefficient

AGENDA

- What is correlation
- Measuring correlation with “the correlation coefficient”
- Determining the level of correlation in a dataset using Pandas
  - Example using Pandas commands on ice-cream data
  - Exercise: determining the level of correlation between variables in the “cars” data set

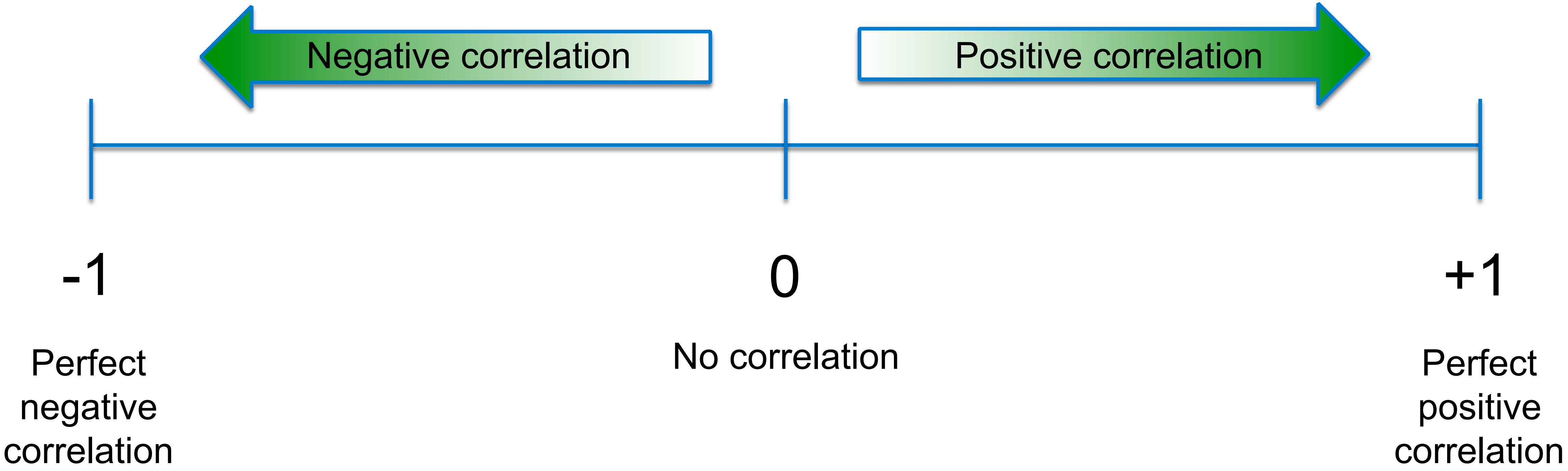
CORRELATION

- **Correlation measures the extent of linear interdependence of two variables**
  - If two variables are correlated, then when the value of one moves the value other tends to also move

- Positively correlated
  - “When the temperature goes up, ice cream sales tend go up”
  - “When ice cream sales go up, the temperature tends to be higher
- Negatively correlated - “When car weight goes up, gas mileage tends to go down”

MEASURING CORRELATION

Pearson's *correlation coefficient* is a commonly used measure of correlation



Negative correlation

Positive correlation

-1

Perfect negative correlation

0

No correlation

+1

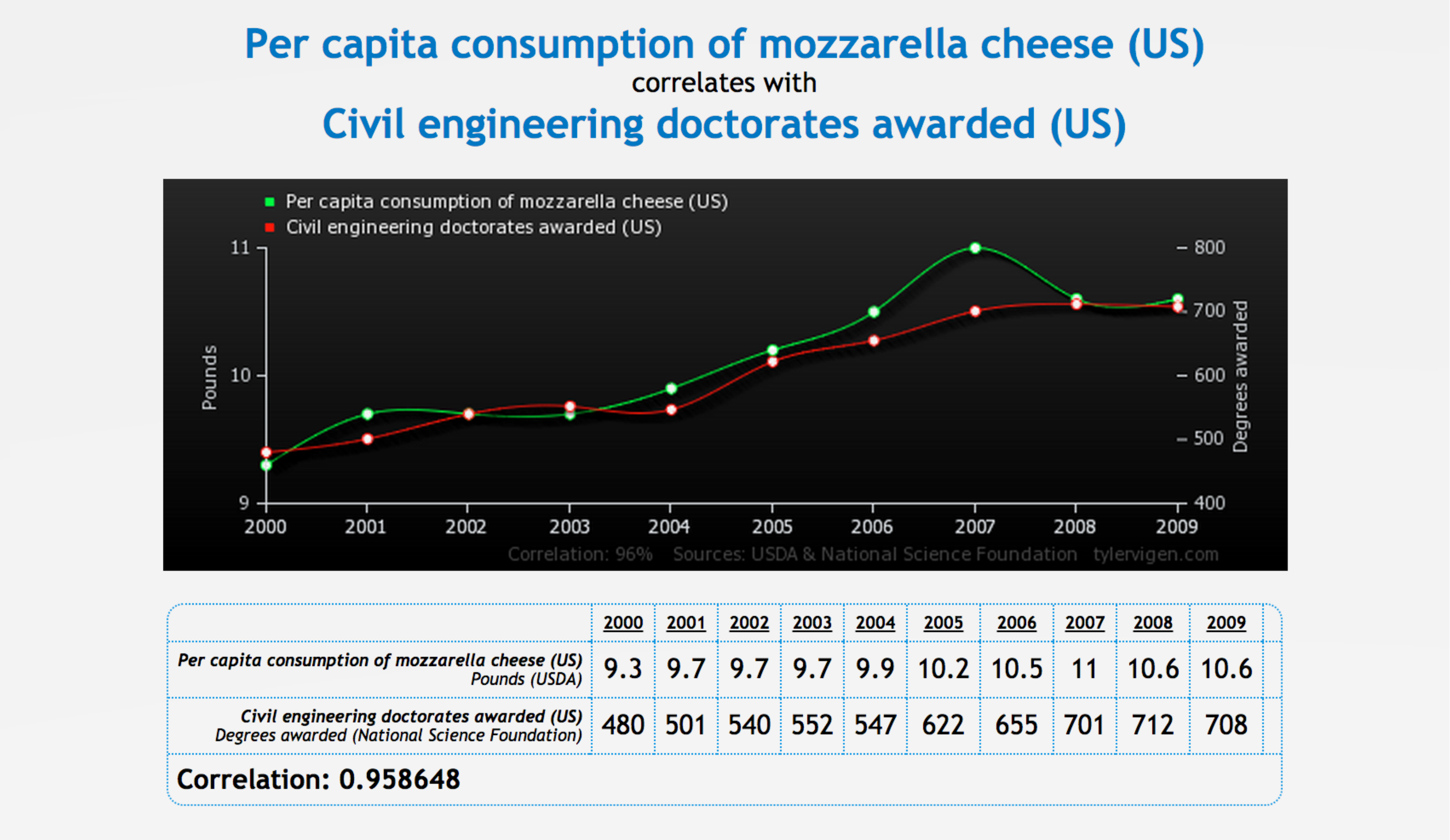
Perfect positive correlation

CORRELATION COEFFICIENT

Positive, negative or no correlation?

- “When the temperature goes up, ice cream sales go up”
- “When beef price rises, steak sales go down”
- Per capita consumption of mozzarella cheese (US), Civil engineering doctorates awarded (US)

SURPRISING CORRELATIONS CAN OCCUR



Per capita consumption of mozzarella cheese (US)

correlates with

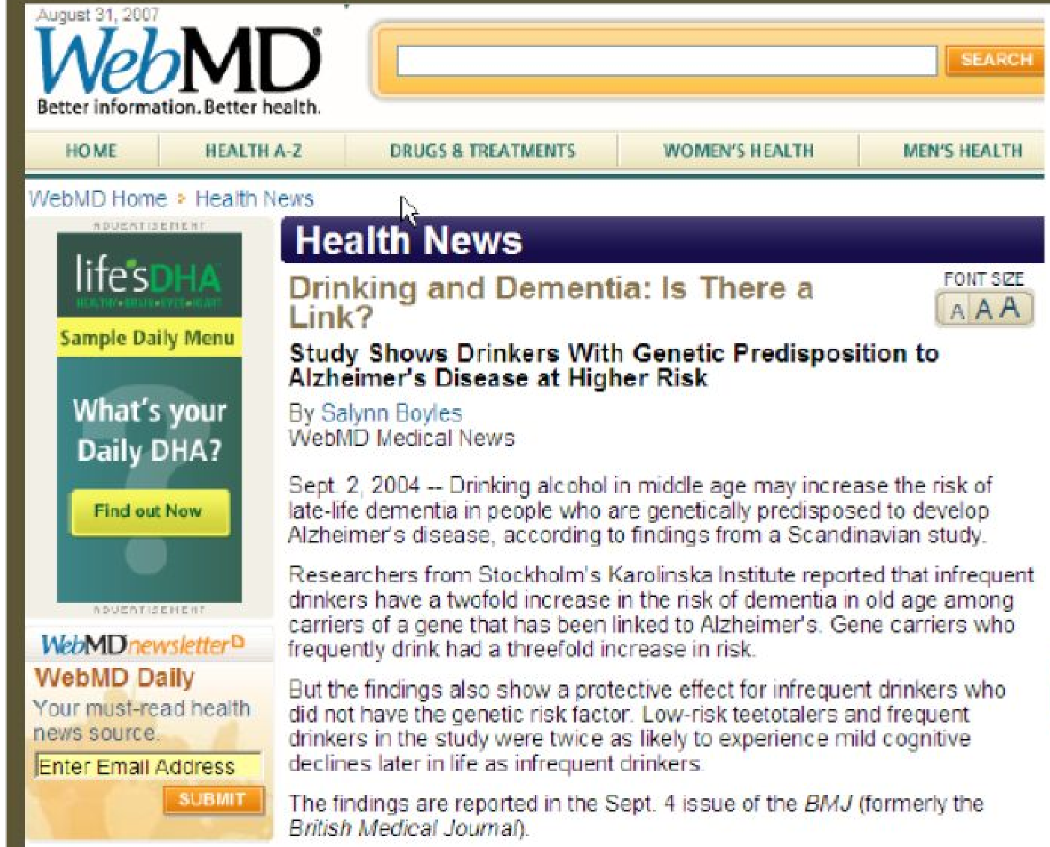
Civil engineering doctorates awarded (US)

🟩 Per capita consumption of mozzarella cheese (US)
🟥 Civil engineering doctorates awarded (US)
Pounds
11
10
9
Degrees awarded
800
700
600
500
400
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 96%
Sources: USDA & National Science Foundation
tylervigen.com

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of mozzarella cheese (US) *Pounds (USDA)*	9.3	9.7	9.7	9.7	9.9	10.2	10.5	11	10.6	10.6
Civil engineering doctorates awarded (US) *Degrees awarded (National Science Foundation)*	480	501	540	552	547	622	655	701	712	708

Correlation: 0.958648

Source: Spurious Correlations, http://tylervigen.com/spurious-correlations



Health News

Drinking and Dementia: Is There a Link?

Study Shows Drinkers With Genetic Predisposition to Alzheimer's Disease at Higher Risk

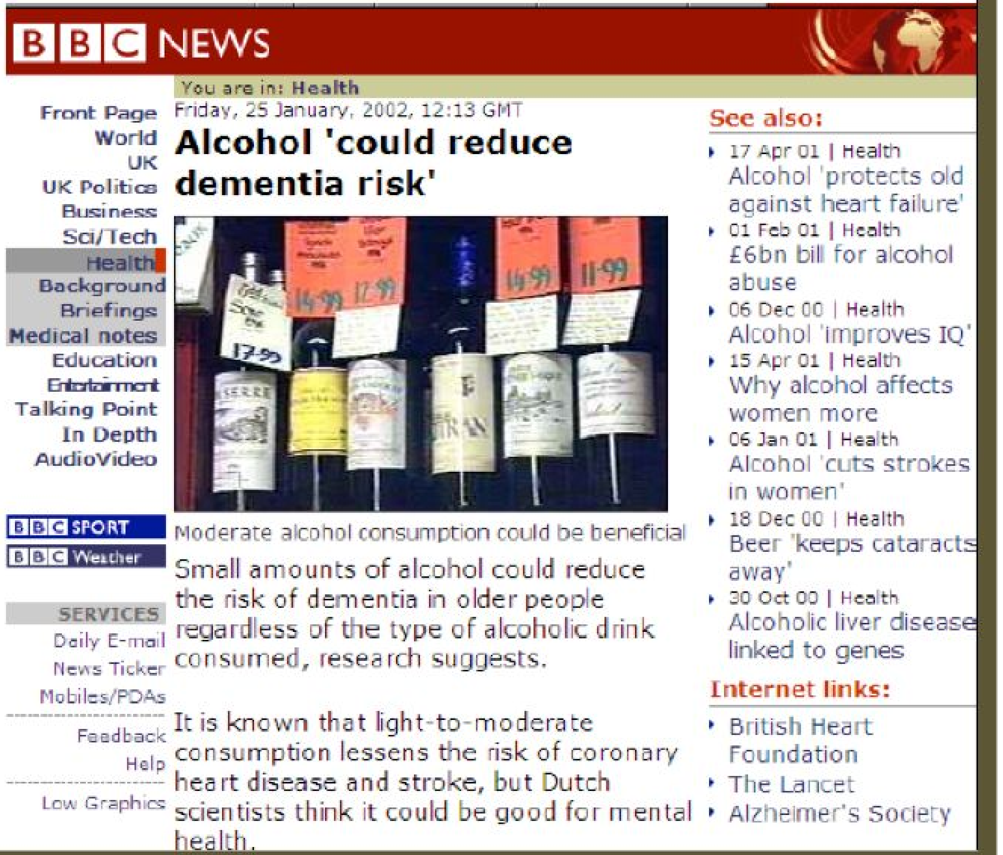
By Salynn Boyles
WebMD Medical News

Sept. 2, 2004 -- Drinking alcohol in middle age may increase the risk of late-life dementia in people who are genetically predisposed to develop Alzheimer's disease, according to findings from a Scandinavian study.

Researchers from Stockholm's Karolinska Institute reported that infrequent drinkers have a twofold increase in the risk of dementia in old age among carriers of a gene that has been linked to Alzheimer's. Gene carriers who frequently drink had a threefold increase in risk.

But the findings also show a protective effect for infrequent drinkers who did not have the genetic risk factor. Low-risk teetotalers and frequent drinkers in the study were twice as likely to experience mild cognitive declines later in life as infrequent drinkers.

The findings are reported in the Sept. 4 issue of the *BMJ* (formerly the *British Medical Journal*).



Friday, 25 January, 2002, 12:13 GMT

Alcohol 'could reduce dementia risk'

Moderate alcohol consumption could be beneficial

Small amounts of alcohol could reduce the risk of dementia in older people regardless of the type of alcoholic drink consumed, research suggests.

It is known that light-to-moderate consumption lessens the risk of coronary heart disease and stroke, but Dutch scientists think it could be good for mental health.

See also:

- 17 Apr 01 | Health
  Alcohol 'protects old against heart failure'
- 01 Feb 01 | Health
  £6bn bill for alcohol abuse
- 06 Dec 00 | Health
  Alcohol 'improves IQ'
- 15 Apr 01 | Health
  Why alcohol affects women more
- 06 Jan 01 | Health
  Alcohol 'cuts strokes in women'
- 18 Dec 00 | Health
  Beer 'keeps cataracts away'
- 30 Oct 00 | Health
  Alcoholic liver disease linked to genes

Internet links:

- British Heart Foundation
- The Lancet
- Alzheimer's Society

CORRELATION DOES NOT IMPLY CAUSATION!

- We *cannot* tell from correlation that there is a cause and effect relationship between two variables

  - Example: A study provides data where health and mood are correlated
    - but improved mood could cause better health, or better heath may cause better mood, they both could be caused by a third factor, or it is just coincidence

- However, a strong correlation can inform us that there *may* be a cause and effect relationship between two variables

CORRELATION ONLY MEASURES THE LINEAR RELATIONSHIP

- It may not reveal relationships between variables that are non-linear

- https://stt.msu.edu/Academics/ClassPages/uploads/SS16/231-1/Summary%20Linear%20Models.pdf

USING PANDAS TO EVALUATE CORRELATION

- Example using the icecream data set
- List data
- Plot data
- Calculate correlation coefficients in a correlation matrix

- Exercise using the built in “car” data set of speeds and stopping distances

TODAY WE LEARNED

- That correlation measures the extent of interdependence of two variables
- How to measuring correlation with “the correlation coefficient”
- That correlation does not = cause and effect
- How to determine the level of correlation in a dataset using Pandas

QUESTIONS?