DATA WRANGLING II

MPA 630: Data Science for Public Management

October 4, 2018

Fill out your reading report on Learning Suite

PLAN FOR TODAY

Missing data

Introduction to correlation

2016 elections, food security, and mortality

MISSING DATA

MATH TIME

oh no

$$1+7+4+2+4=?$$

$$1+7+4+?+4=?$$

$$\frac{1+7+4+2+4}{5}=?$$

$$\frac{1+7+4+?+4}{5}=?$$

CODE TIME

```r
numbers <- c(1, 7, 4, 2, 4)
```

```r
sum(numbers)
```

```
[1] 18
```

```r
mean(numbers)
```

```
[1] 3.6
```

```r
numbers <- c(1, 7, 4, NA, 4)
```

```r
sum(numbers)
```

```
[1] NA
```

```r
mean(numbers)
```

```
[1] NA
```

```r
sum(numbers, na.rm = TRUE)
```

```
[1] 16
```

```r
mean(numbers, na.rm = TRUE)
```

```
[1] 4
```

IS IT OKAY TO REMOVE MISSING DATA?

It depends!

REASONS FOR MISSING DATA

MCAR	MAR	MNAR
Missing completely at random	Missing (conditionally) at random	Missing not at random
No relationship between missingness and either observed or unobserved data	Relationship between missingness and **observed** data, but not unobserved data	Relationship between missingness and data

EXAMPLES

MCAR

Researcher accidentally doesn't record some of the values

MAR

Men are less likely to fill out survey about depression, but not because of their depression

(but because maleness makes them less likely to do so)

MNAR

Men are less likely to fill out a survey about depression *because of* their depression

WHAT TO DO

MCAR	MAR	MNAR
Safely ignore!	Mostly safely ignore! Account for the thing that makes it missing (e.g. look at maleness specifically when analyzing depression survey response)	Don't ignore! Explain why it's missing and incorporate that into analysis

0 vs. NA

Sometimes NAs are really 0s

	county	population_2010	religion	adherents	adherents_per_1000	congregations	religion_clean
1	Beaver County	6629	AG	*NA*	*NA*	*NA*	Assemblies of Goc
2	Beaver County	6629	BUDM	*NA*	*NA*	*NA*	Buddhism, Mahay
3	Beaver County	6629	CATH	152	22.9288889	1	Catholic
4	Beaver County	6629	EC	*NA*	*NA*	*NA*	Episcopal Church
5	Beaver County	6629	EVAN	28	4.2188889	2	Evangelical
6	Beaver County	6629	GRK	*NA*	*NA*	*NA*	Greek Orthodox
7	Beaver County	6629	LCMS	*NA*	*NA*	*NA*	Lutheran Church,
8	Beaver County	6629	LDS	4965	748.9800000	15	LDS
9	Beaver County	6629	MPRT	*NA*	*NA*	*NA*	Mainline Protestar
10	Beaver County	6629	MSLM	*NA*	*NA*	*NA*	Muslim

```
mutate(adherents = ifelse(is.na(adherents), 0, adherents))
```

INTRODUCTION TO CORRELATION

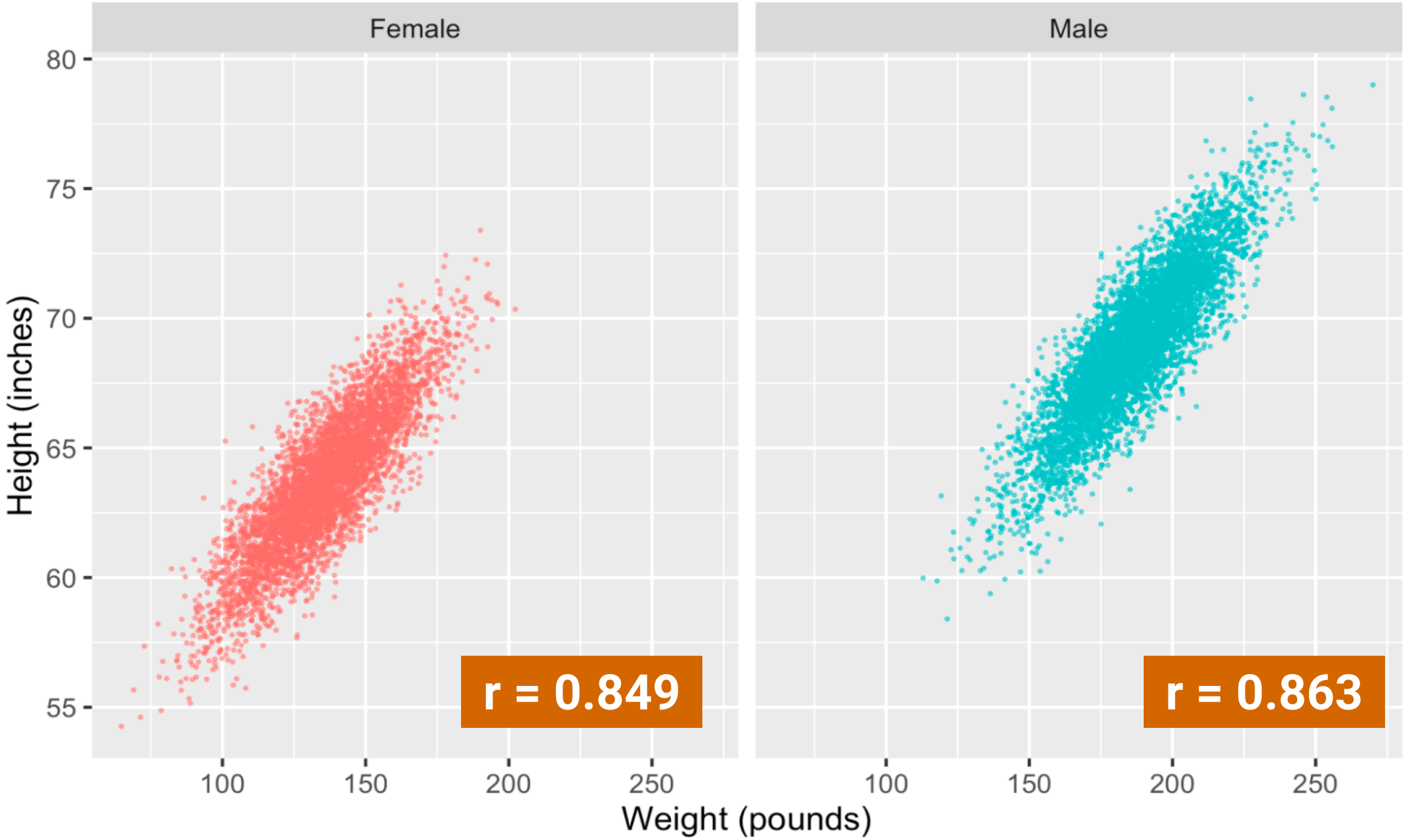
CORRELATION

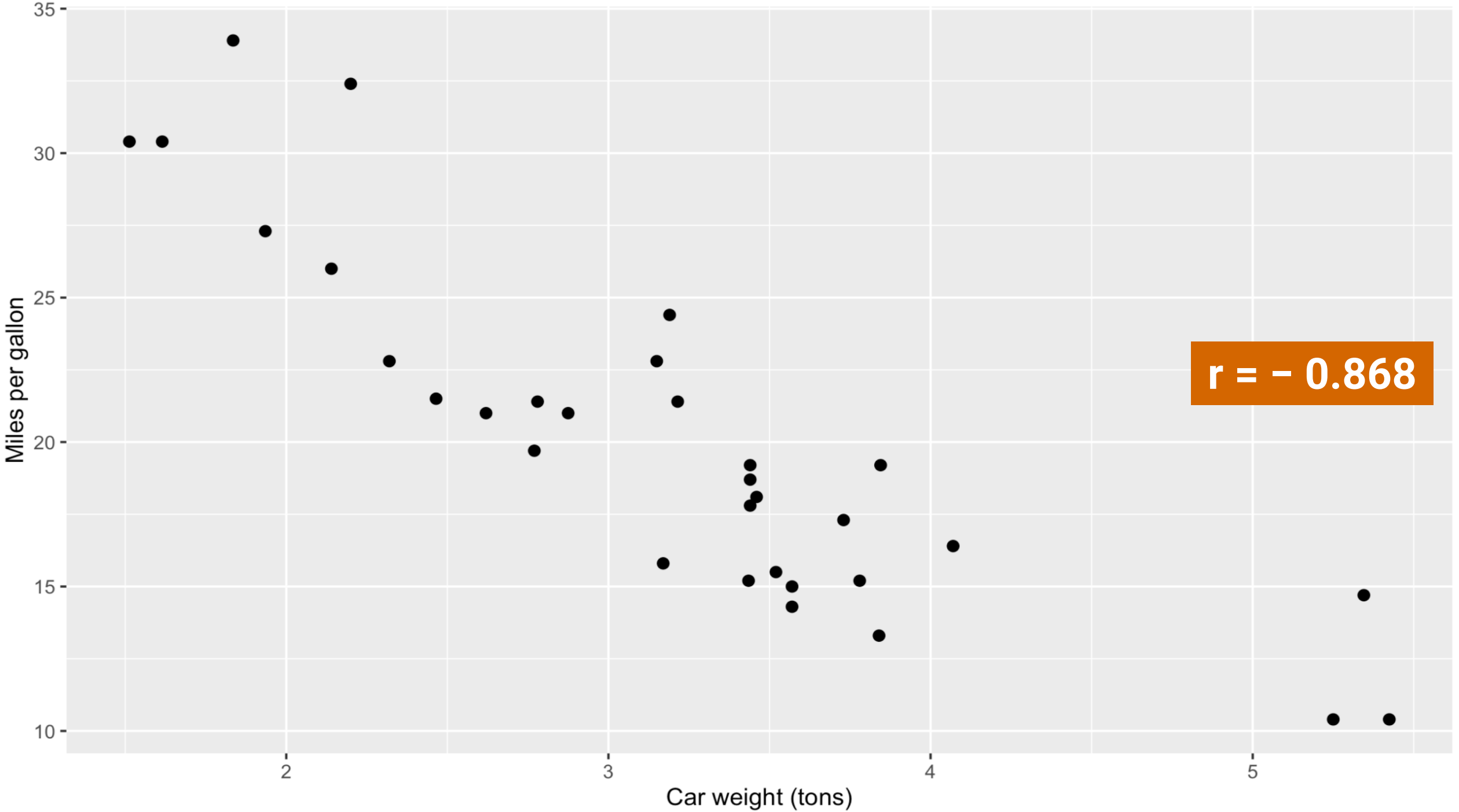
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

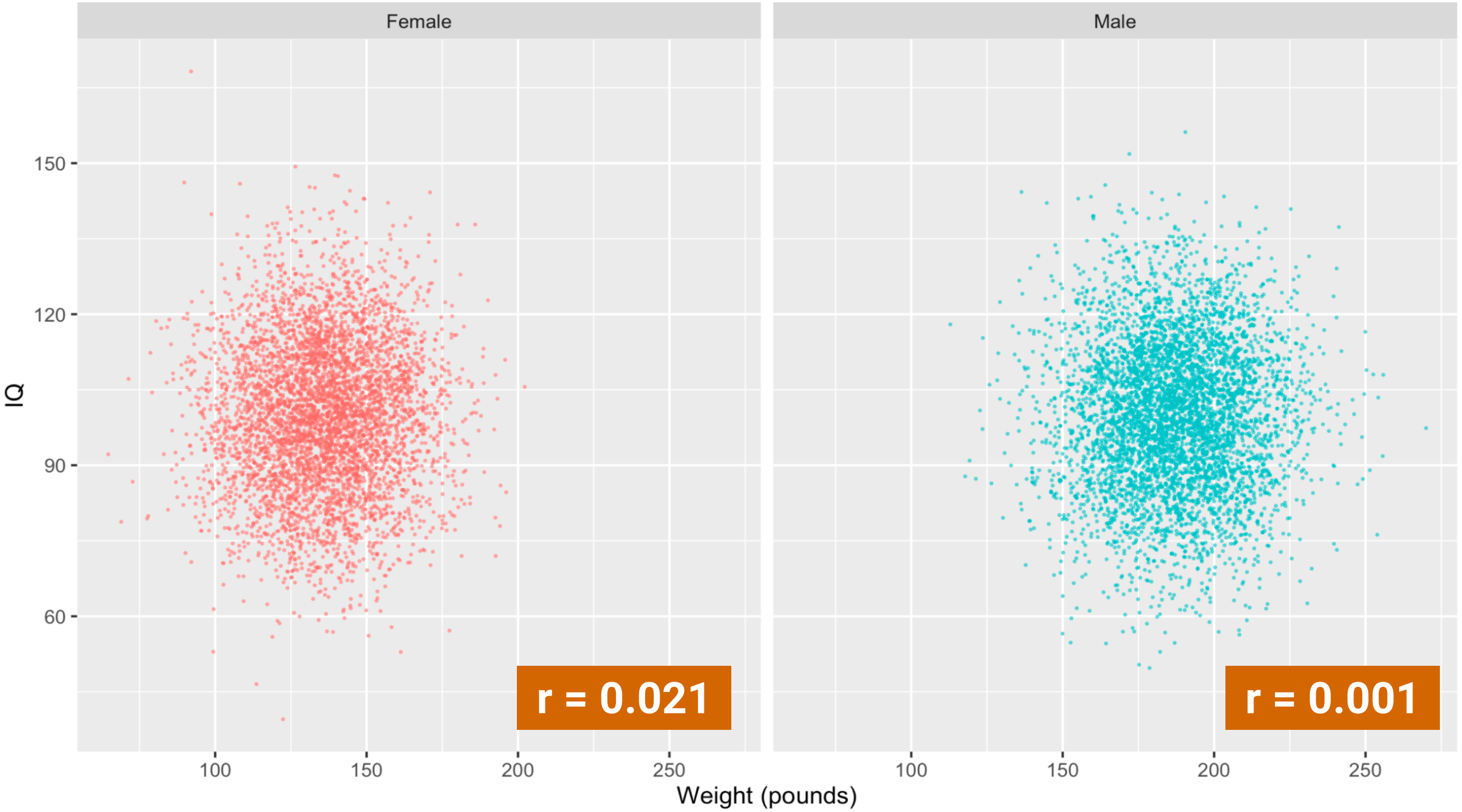
How closely two variables are related + direction of relation

-1 to 1

-1 and 1 = perfectly correlated;
0 = perfectly uncorrelated


Female
Male
Height (inches)
80
75
70
65
60
55
100
150
200
250
Weight (pounds)
r = 0.849
r = 0.863


35
30
25
20
15
10
Miles per gallon
2
3
4
5
Car weight (tons)
r = − 0.868


Female
Male
IQ
150
120
90
60
100
150
200
250
Weight (pounds)
r = 0.021
r = 0.001

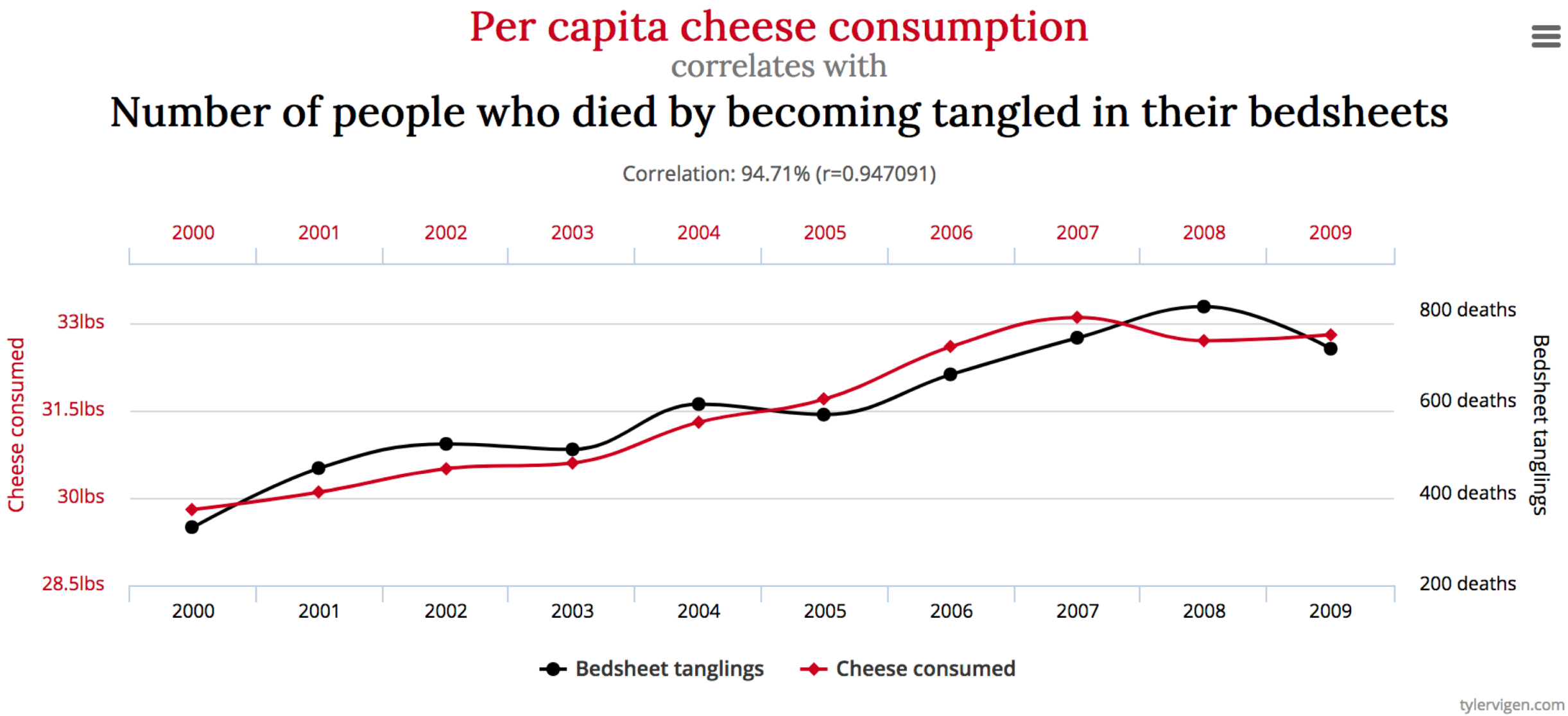
GENERAL GUIDELINES

0	No relationship
0.01–0.19	Little to no relationship
0.20–0.29	Weak relationship
0.30–0.39	Moderate relationship
0.40–0.69	Strong relationship
0.70–0.99	Very strong relationship
1	Perfect relationship

Can be positive or negative

BEWARE OF C E

Spurious correlations



2016 ELECTIONS, FOOD SECURITY, AND MORTALITY

Vince Feula
Dallen Done
Aleni Regehr
Mike Hall
Stephanie Livsey

Grant Gillum
Carina Alleman
Tim Anderson
Krista Gardner

Karla Ward
Sione Manoa
Brad Lester
Wallis Rothlisberger

Angie Anderson
Cindy Tolman
Daniel Dudley
Spencer Foster

Brian Tuttle
Joel Cook
Natalie Gay

Craig Haderlie
Rich Christianson
Phyllis Nielsen
Sara Donakey

Teri Chatterton
Lissa Camacho
Jeff Long
Julie Rash

Michelle Stevenson
Colleen Kohler
Spencer Parkinson
Elyse Bradley

Rebecca Smoot
Mark Gefrom
Hilary O'Neil
Harvey Unga
Dave Hawks

Dallas Reynolds
Dane Larsen
Tracy McIntier
Heidi Hatch