Quantitative methods

Week #8-9

Gergely Daróczi

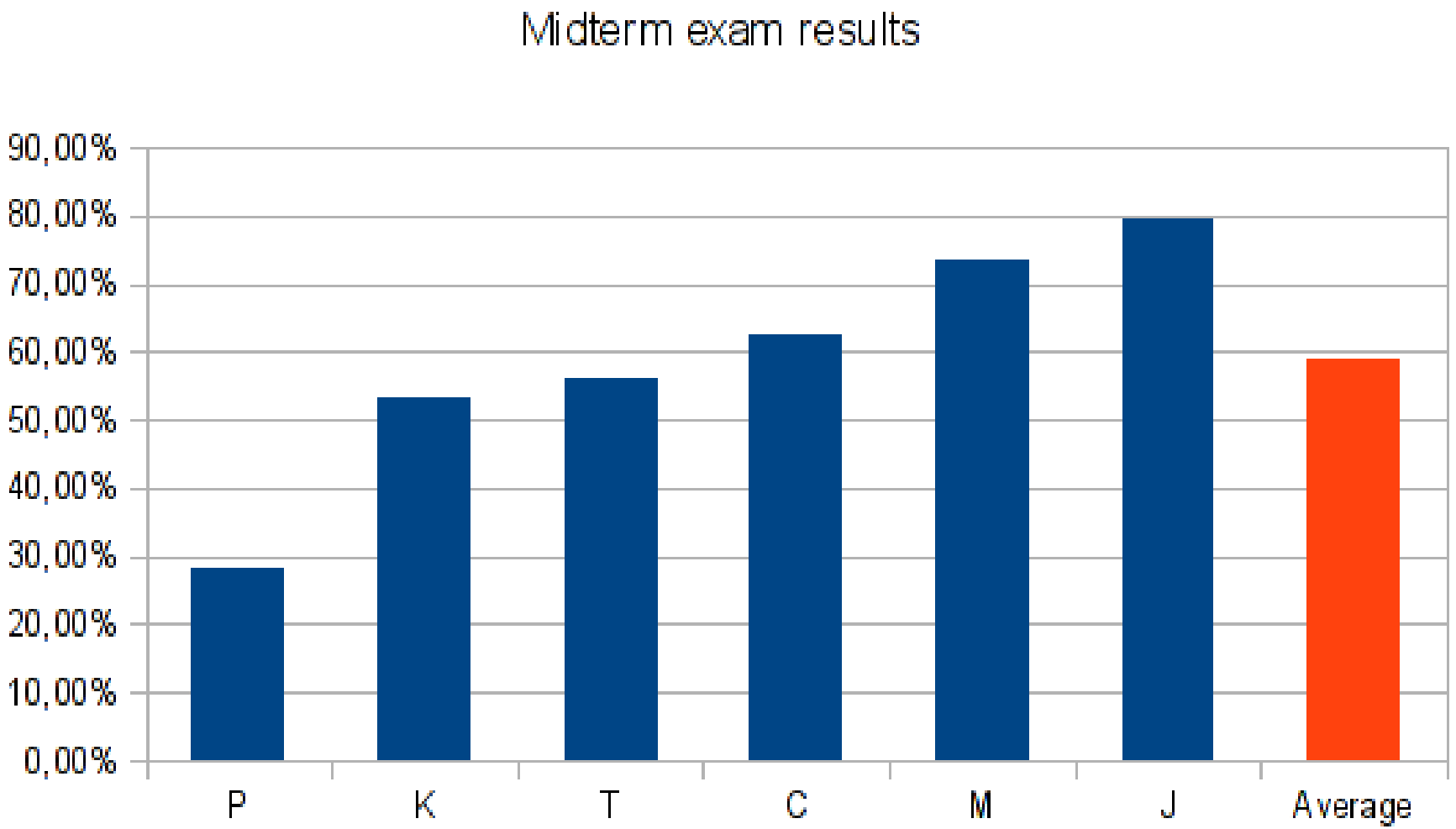
Corvinus University of Budapest, Hungary

5 April 2013

Outline

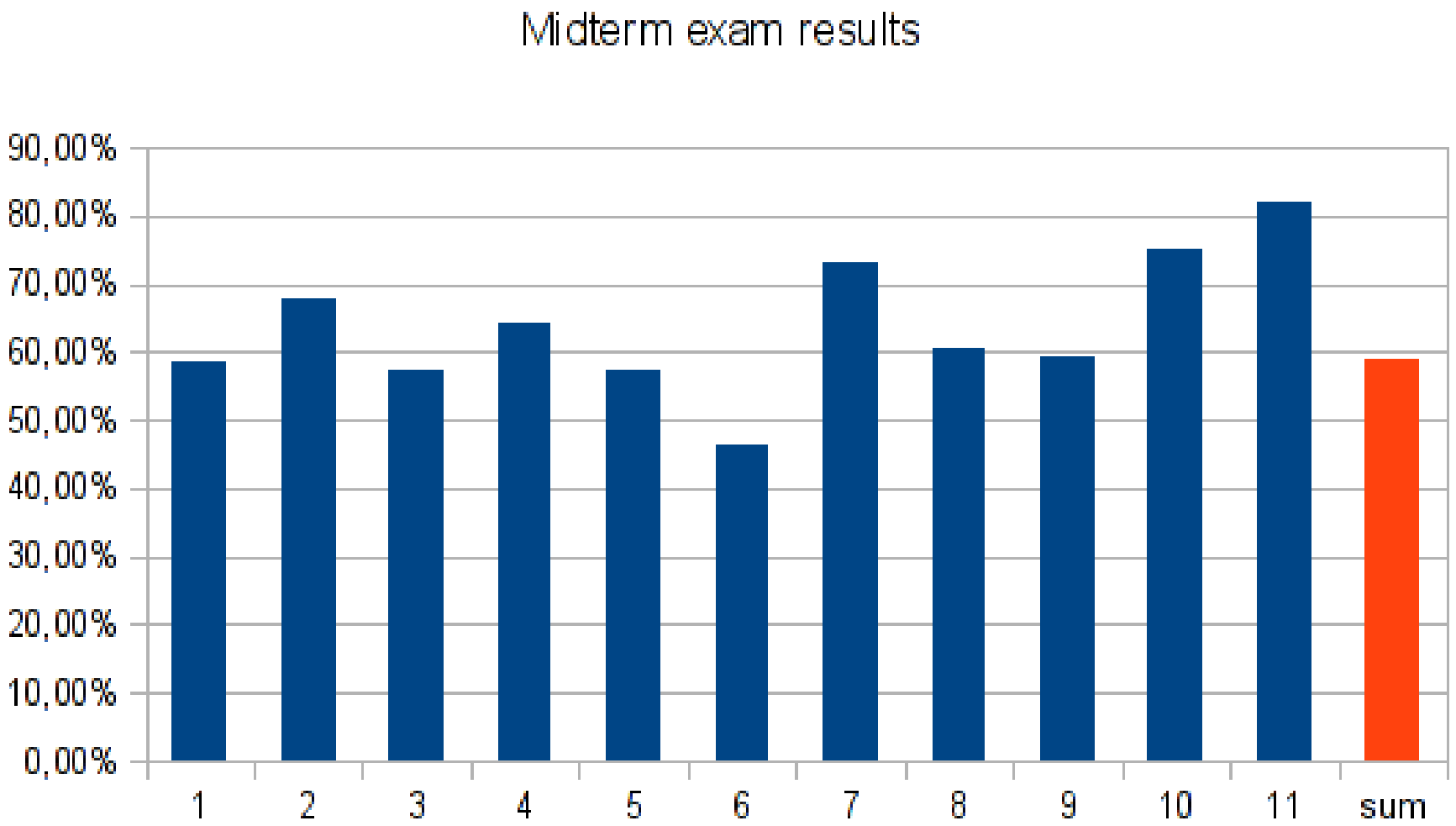
Midterm exam

Results



Midterm exam

Results



Final examination questions

Comprehensive exam

Singleton, R. A. Jr. and Bruce C. Straits (1999): Approaches to Social Research. Third Edition. Oxford University Press: New York/Oxford.

Questions:

1. What is reliability? How do the main rules concerning the order of survey questions improve the reliability and validity of survey data? (pp. 113-117, 292-296)
2. What is meant by probability sampling? How do stratification and multistage cluster sampling affect sampling errors? Why? (pp. 141-142, 145-156)
3. What are the main types of non-probability sampling? Explain why these types do not meet the criteria of probability samples. (pp. 157-169)
4. What factors affect the desired sample size? (pp. 163-169)

Levels of Measurement

Qualitative and quantitative variables in depth

Qualitative variables:

- Nominal: exhaustive labels with no intersect (mutual exlcusivity) not in a specific order
- Ordinal: an (possible) ordered variable with exhaustive labels not intersecting

	Nominal	Ordinal	Interval	Ratio
Classification	X	X	X	X
Rank order		X	X	X
Equal intervals			X	X
Nonarbitrary zero				X

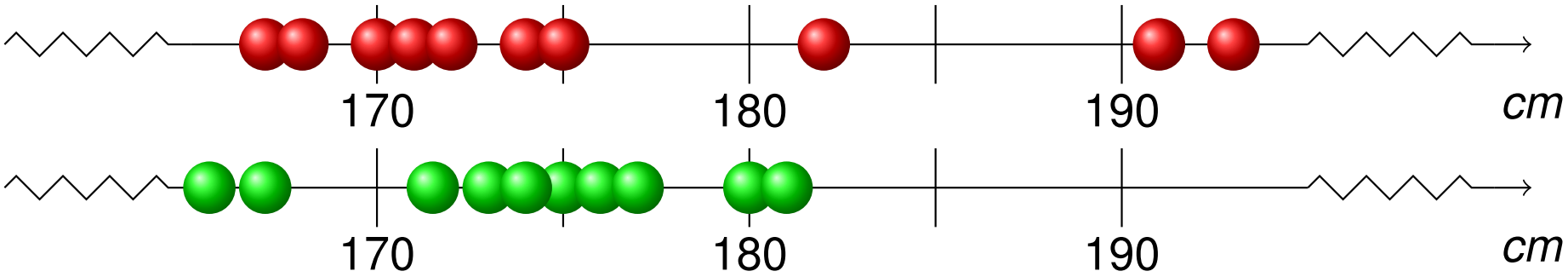
Quantitative variables:

- Interval: equal distances between the ordered labels (numbers)
- Ratio: a scale with a zero point

For Simple Random Sampling:

- mean: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- median
- mode
- min, max, range
- standard deviation: $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$
- variance
- IQR

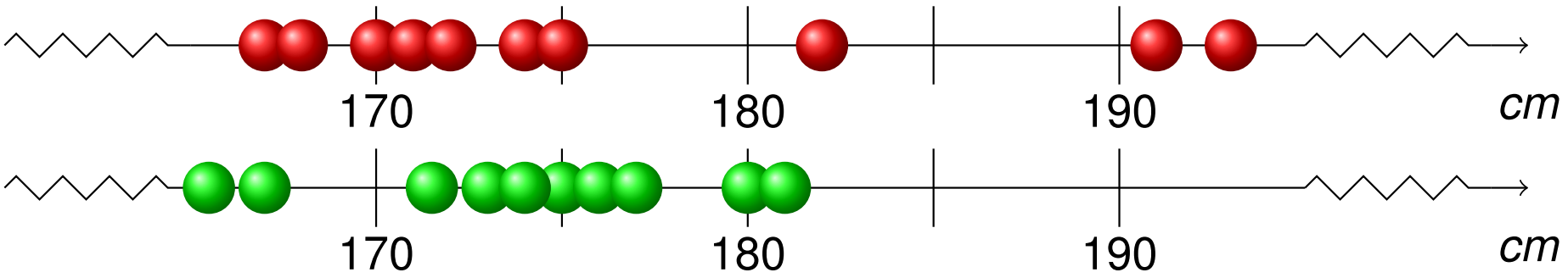
We have measured 10-10 students in two classrooms.

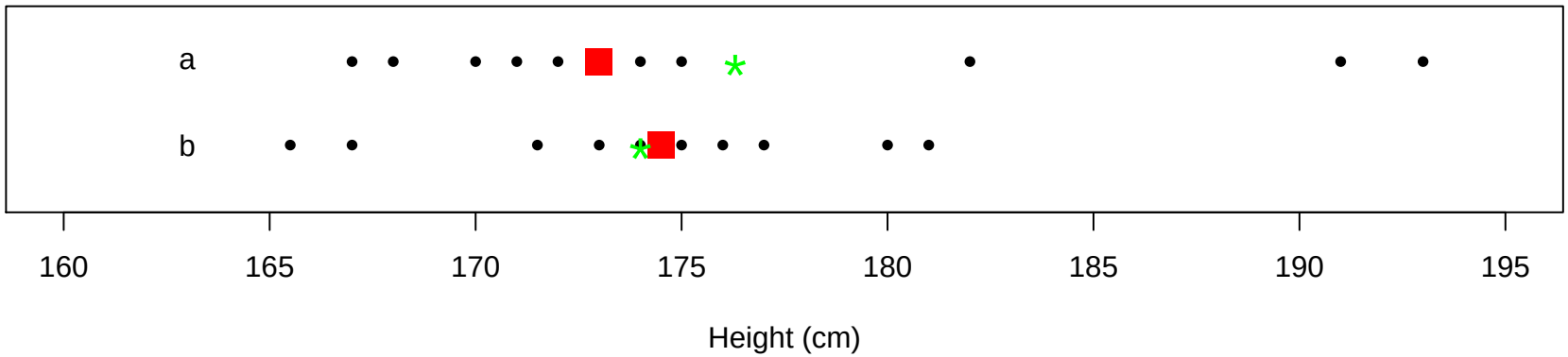


Which class has higher students based on this small sample? Think about averages as good estimates of population parameters!

Descriptive statistics

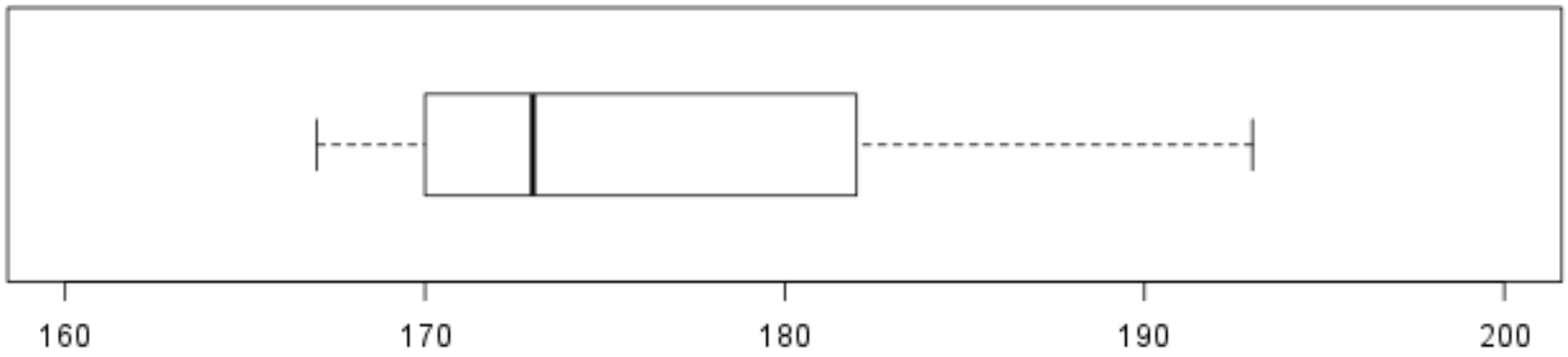
Averages

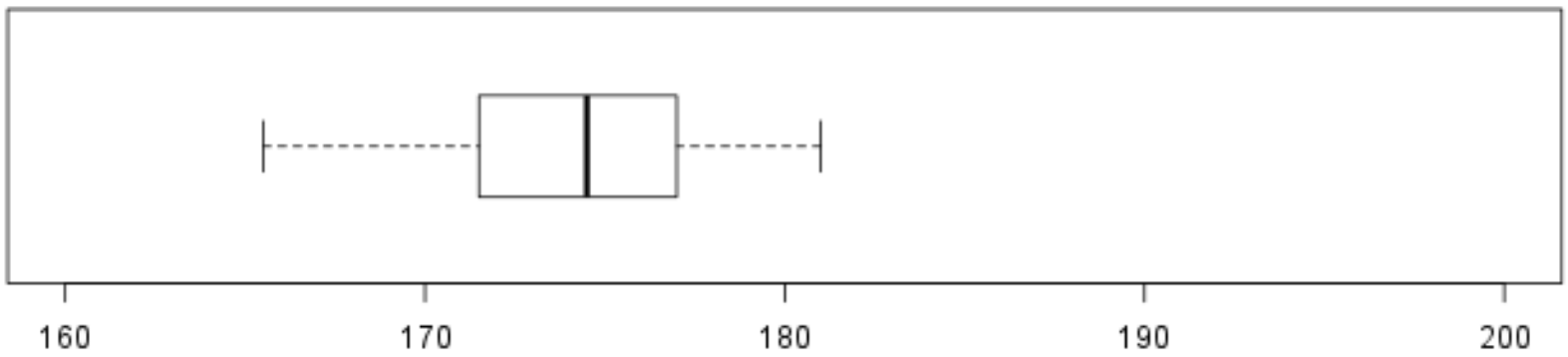




Descriptive statistics

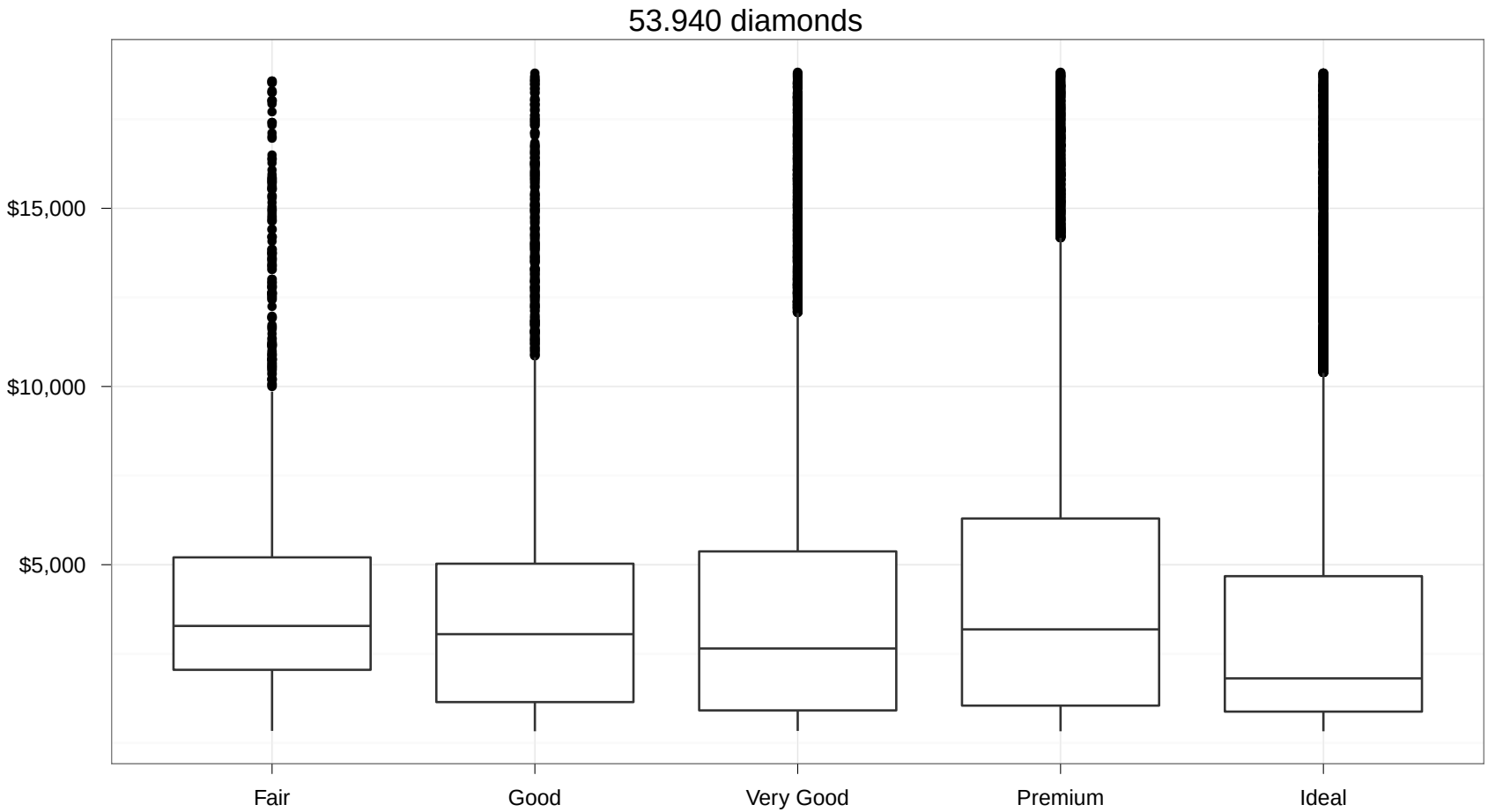
Median and IQR





Relation between variables

A visual example




```
ggplot(diamonds, aes(cut, price)) + geom_boxplot() + xlab('') + ylab('') +
scale_y_continuous(formatter="dollar") + theme_bw() + opts(title="53.940 diamonds")
```

The structure of the demo dataset

ggplot2/diamonds

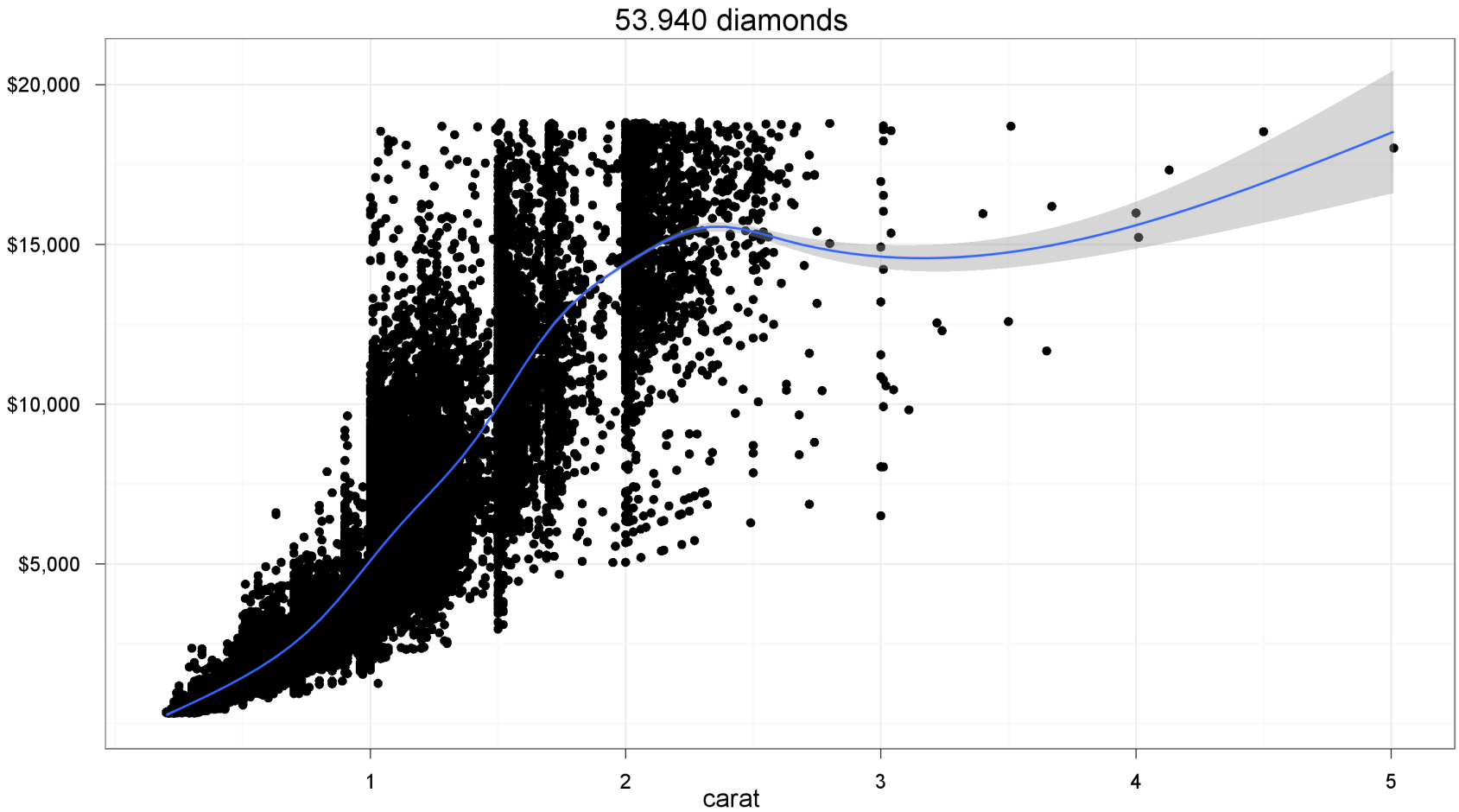
```
Prices of 50,000 round cut diamonds

Description:
     A dataset containing the prices and other attributes of almost
     54,000 diamonds.  The variables are as follows:
```

- price. price in US dollars (\$326--\$18,823)
- carat. weight of the diamond (0.2--5.01)
- cut. quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- colour. diamond colour, from J (worst) to D (best)
- clarity. a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
- x. length in mm (0--10.74)
- y. width in mm (0--58.9)
- z. depth in mm (0--31.8)
- depth. total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
- table. width of top of diamond relative to widest point (43--95)

Relation between variables

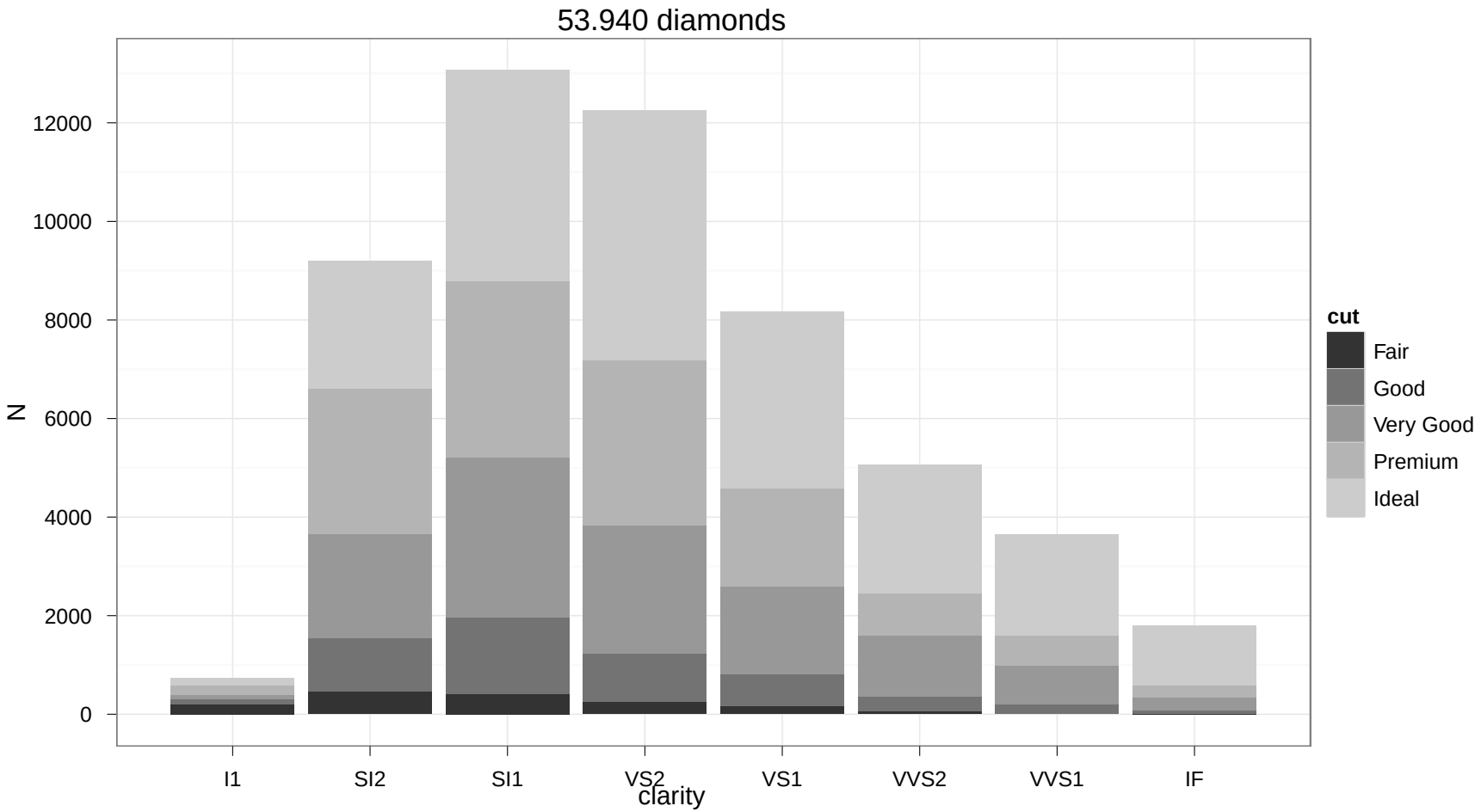
A visual example




```
ggplot(diamonds, aes(carat, price)) + geom_point() + geom_smooth() + ylab('') +
scale_y_continuous(formatter="dollar") + theme_bw() + opts(title="53.940 diamonds")
```

Relation between variables

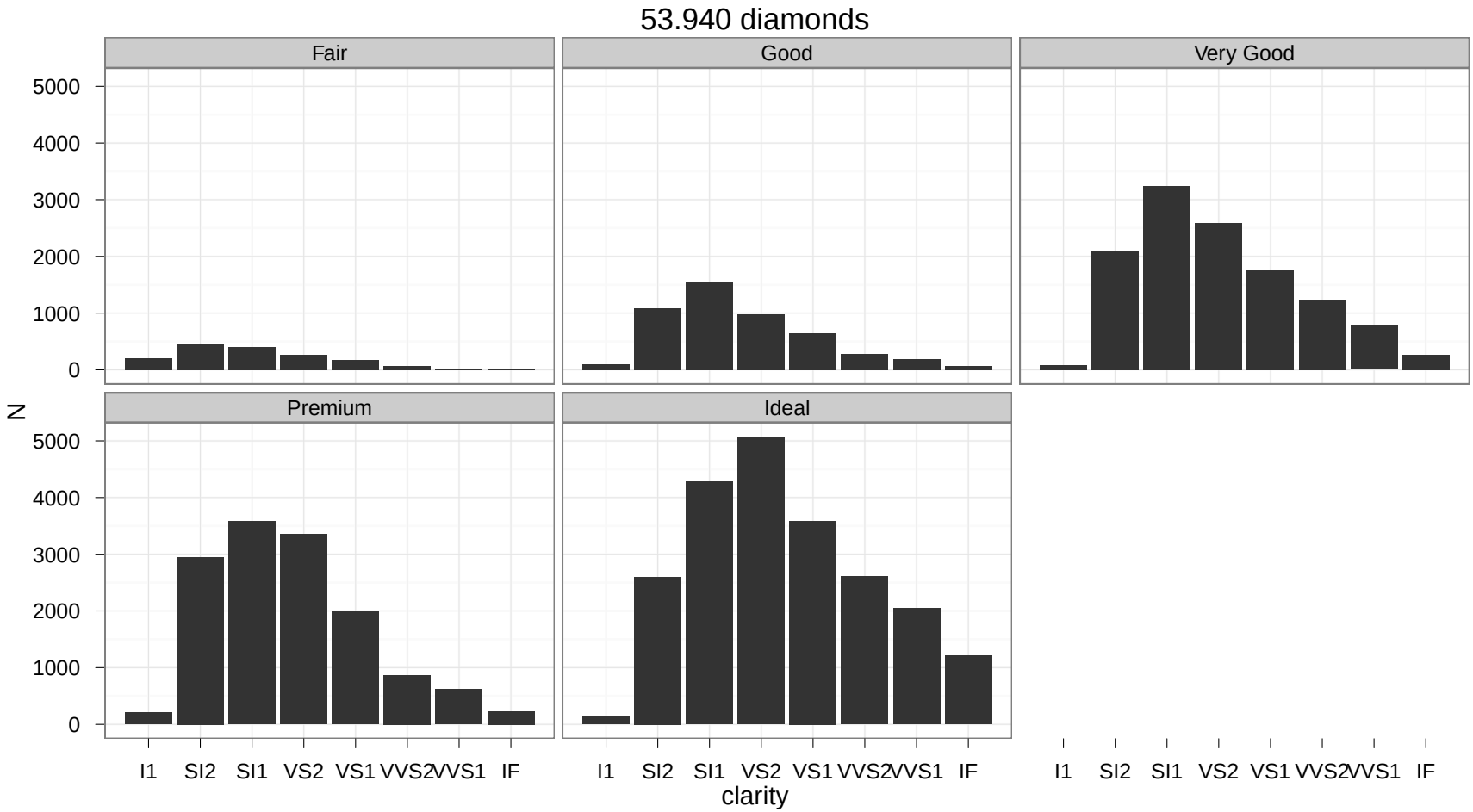
A visual example




```
ggplot(diamonds, aes(clarity, fill=cut)) + geom_bar() + ylab("N") +
theme_bw() + opts(title="53.940 diamonds")
```

Relation between variables

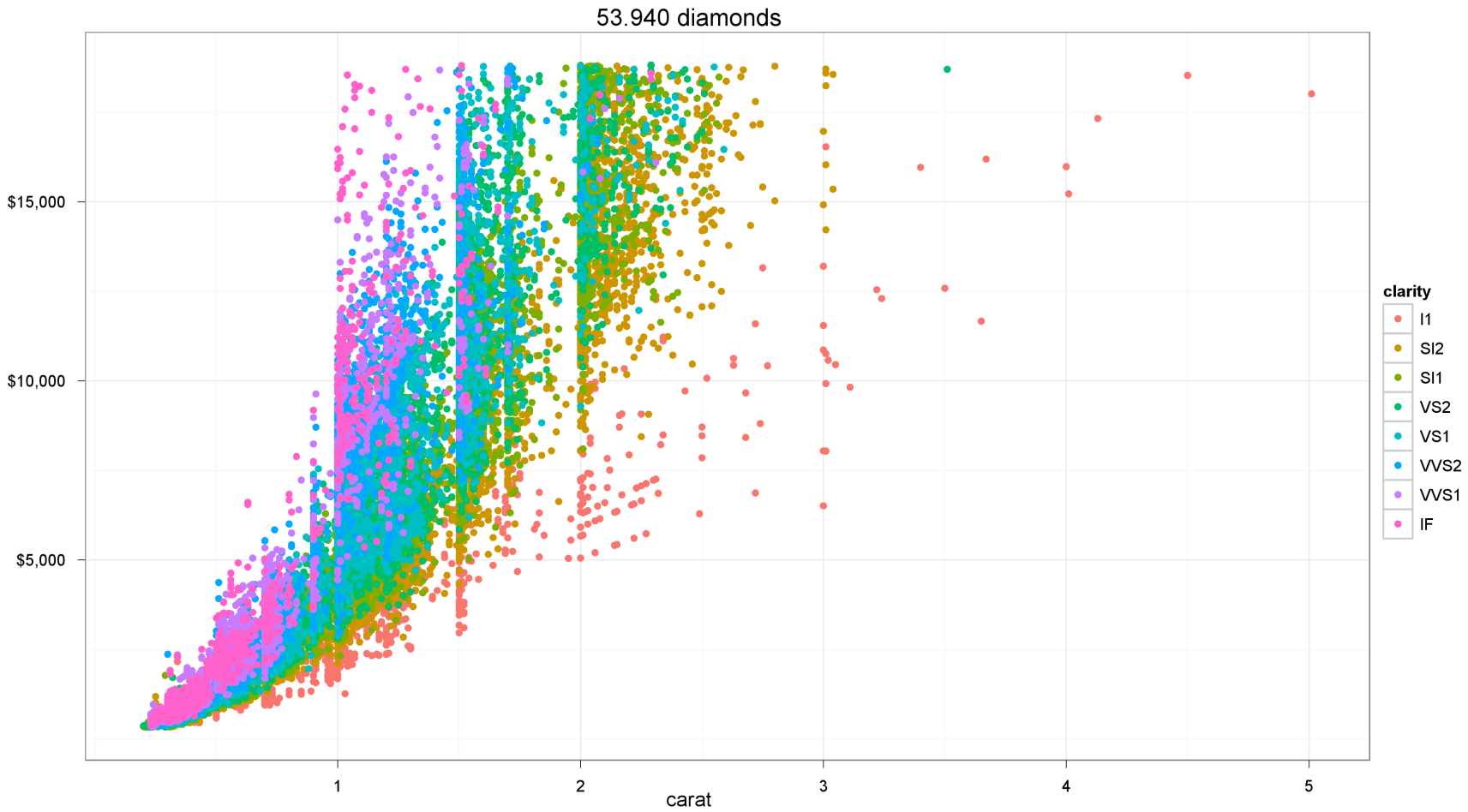
A visual example




```
ggplot(diamonds, aes(clarity)) + geom_bar() + ylab("N") + facet_wrap(~ cut) +
theme_bw() + opts(title="53.940 diamonds")
```

Relation between variables

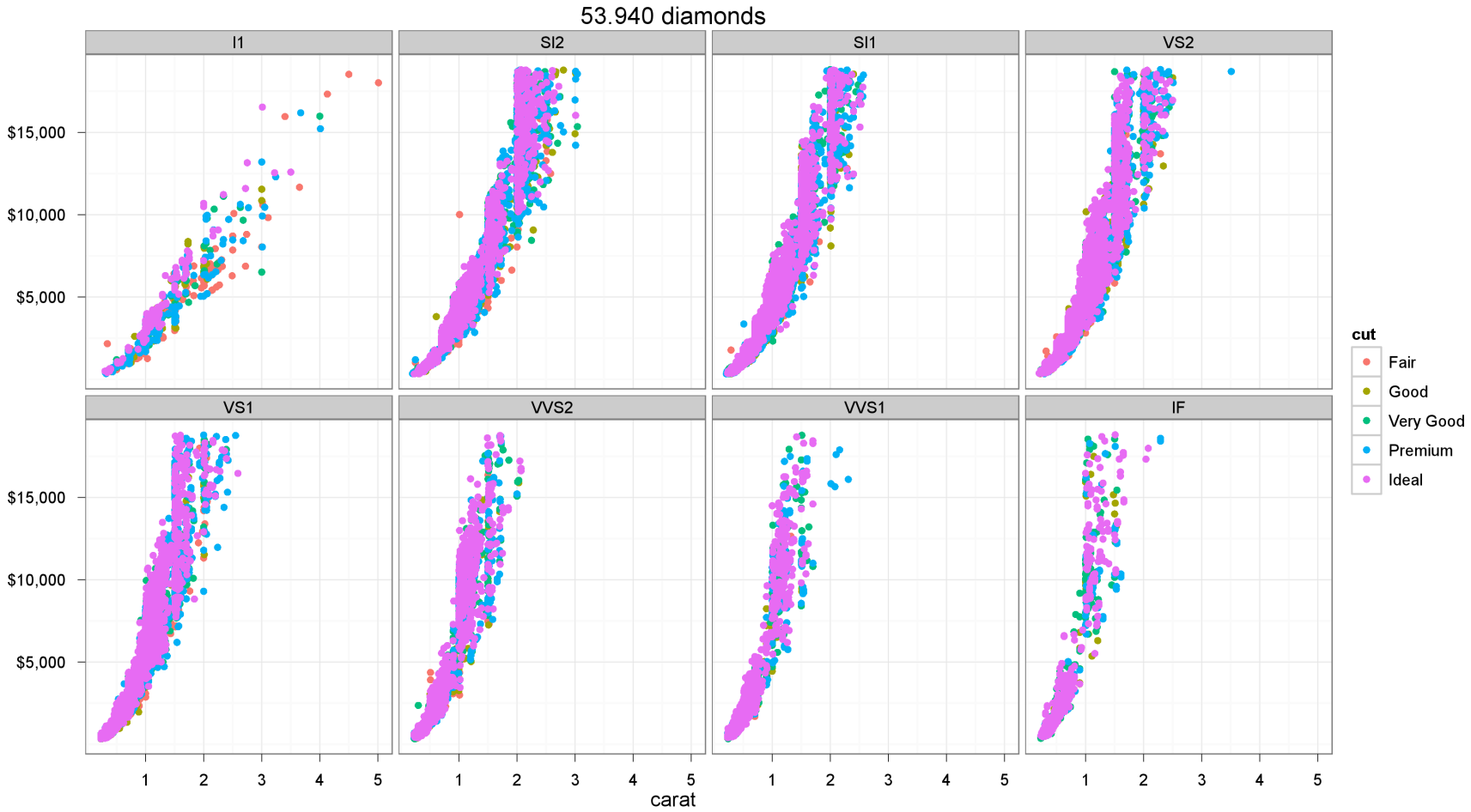
A visual example




```
ggplot(diamonds, aes(carat, price, color=clarity)) + geom_point() + ylab('') +
scale_y_continuous(formatter="dollar") + theme_bw() + opts(title="53.940 diamonds")
```

Relation between variables

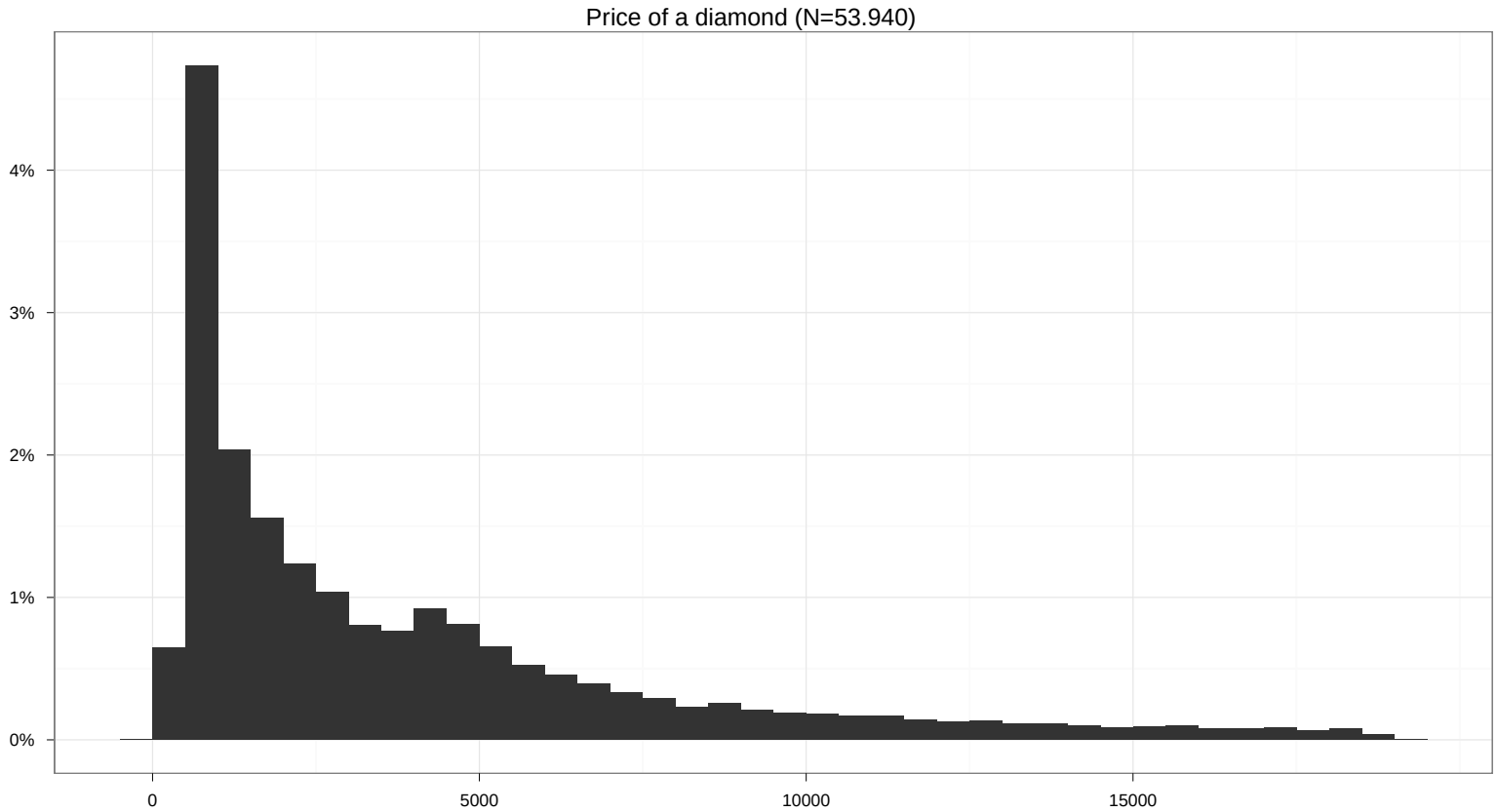
A visual example




```
ggplot(diamonds, aes(carat, price, color=cut)) + geom_point() + ylab('') + facet_wrap(~ clarity,nr
scale_y_continuous(formatter="dollar") + theme_bw() + opts(title="53.940 diamonds")
```

Test your knowledge!

Reliability and validity



A survey was taken place about diamonds available for sale on the Internet.

What do you think of the reliability and validity of this research?

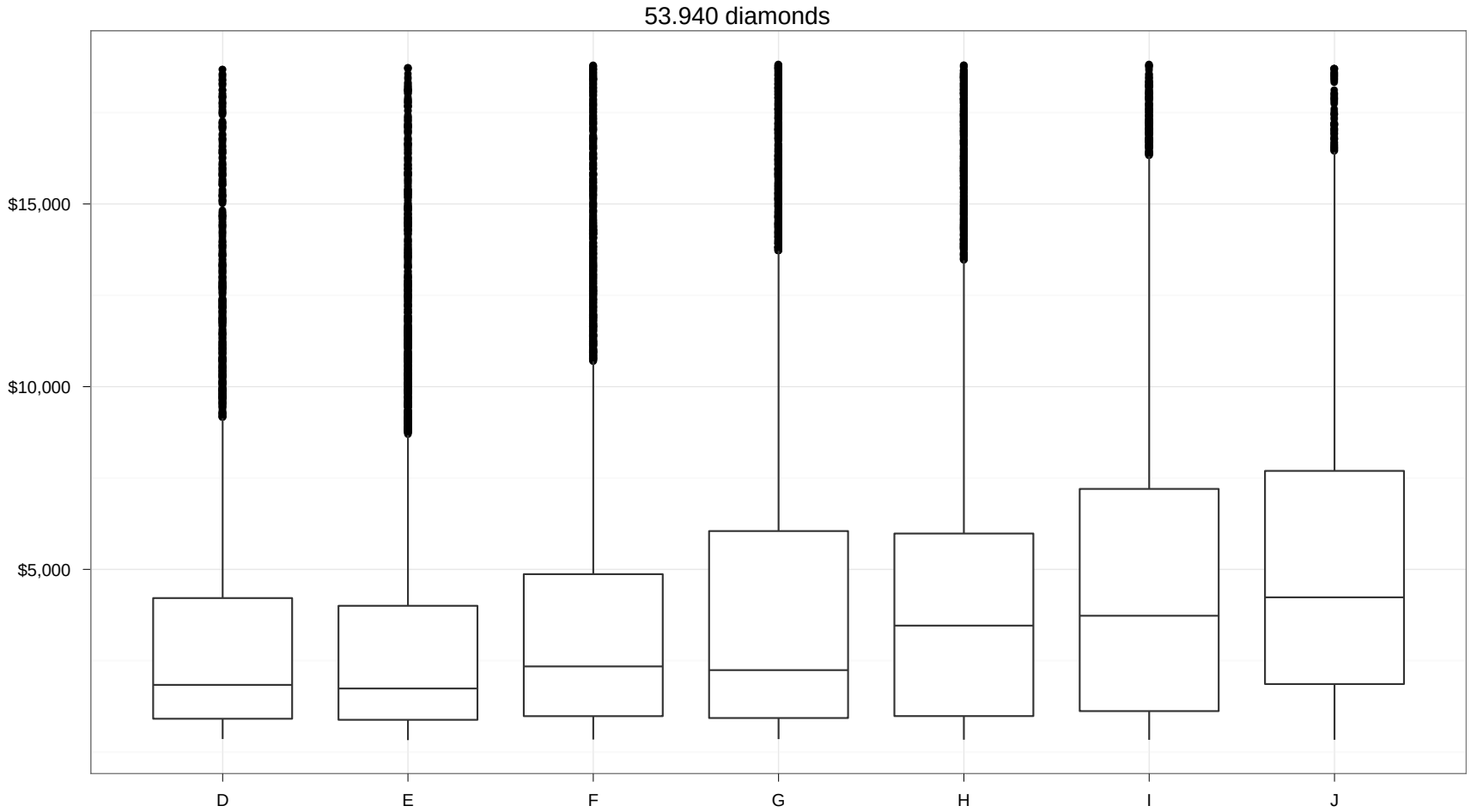
Possible relationship between variables:

- association,
- correlation,
- spurious relationship,
- influence,
- direction of influence,
- **causality**.

STATISTICALLY SIGNIFICANT

Relation between variables

Direction of influence




```
ggplot(diamonds, aes(color, price)) + geom_boxplot() + xlab('') + ylab('') +
scale_y_continuous(formatter="dollar") + theme_bw() + opts(title="53.940 diamonds")
```

Cheatsheat

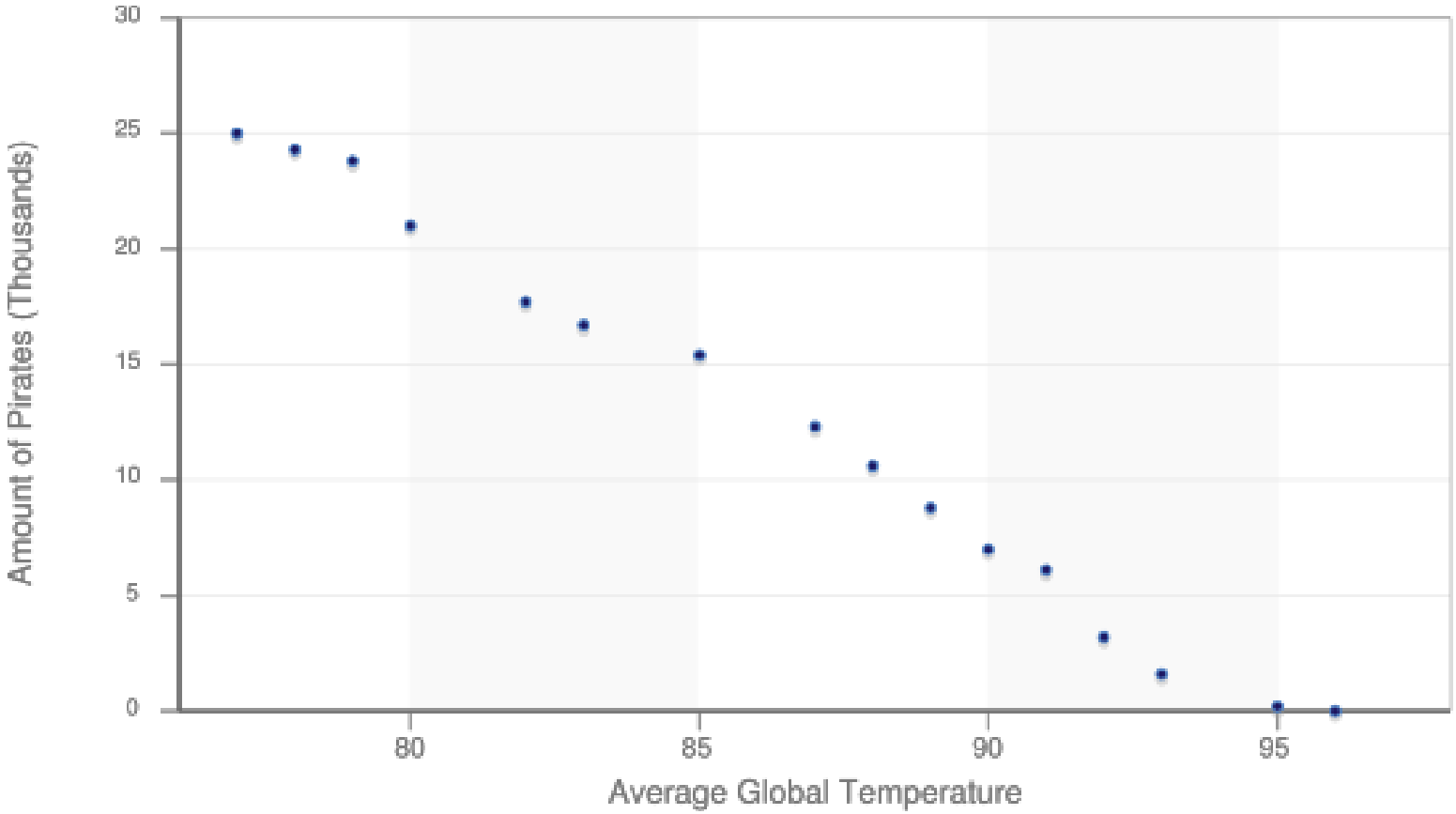
Diamonds' colors

GIA	Status: *current*	AGS	Status: *current*	AGS	Status: *historical: pre 1995*	CIBJO	Status: *current*	IDC	Status: *current*	Scan. D.N.	Status: *current*	Old World Terms	Status: *historical*
grade and description[6]		*grade and electronic colorimeter scale*[7]		*grade and electronic colorimeter scale*[7]		*grade*[8]		*grade and description*[8]		*grade for .50ct and over*[9]	*grade for under .50ct*	*series 1 scale*[8]	*series 2 scale*[8]
D	*Colorless*	0	0–0.49	0	0–0.75	Exceptional white +	Exceptional white +	*Colorless*		River	White	Finest White	Jager
E		0.5	0.5–0.99	1	0.76–1.35	Exceptional white	Exceptional white						River
F		1.0	1.0–1.49	2	1.36–2.00	Rare white +	Rare white +	*Colorless when viewed through the crown*		Top Wesselton		Fine White	
G	*Near Colorless*	1.5	1.5–1.99			Rare white	Rare white						Top Wesselton
H		2.0	2.0–2.49	3	2.01–2.50	White	White			Wesselton		White	Wesselton
I		2.5	2.5–2.99	4	2.51–3.0	Slightly tinted white	Slightly tinted white	*Slightly colored*		Top Crystal	Slightly tinted white	Commercial White	Top Crystal
J		3.0	3.0–3.49	5	3.01–3.75					Crystal		Top silver cape	Crystal
K	*Faint Yellow*	3.5	3.5–3.99	6	3.76–4.5	Tinted white	Tinted white			Top cape	Tinted white	Silver cape	Top cape
L		4.0	4.0–4.49										
M		4.5	4.5–4.99	7	4.51–5.50	Tinted color 1	Tinted color	*Slightly colored to colored*		Cape	Tinted color	Light cape	Cape
N	*Very Light Yellow*	5.0	5.0–5.49			Tinted color 2							Low Cape
O		5.5	5.5–5.99	8	5.51–7.0					Light yellow		Cape	Very light yellow
P		6.0	6.0–6.49										Light yellow
Q		6.5	6.5–6.99										
R		7.0	7.0–7.49	9	7.01–8.5							Dark cape	
S	*Light Yellow*	7.5	7.5–7.99			Tinted color 3				Yellow			
T		8.0	8.0–8.49										
U		8.5	8.5–8.99	10	8.51–10								
V		9.0	9.0–9.49										
W		9.5	9.5–9.99										
X		10	10+	10+									
Y													
Z													

Source: http://en.wikipedia.org/wiki/Diamond_color

Relation between variables

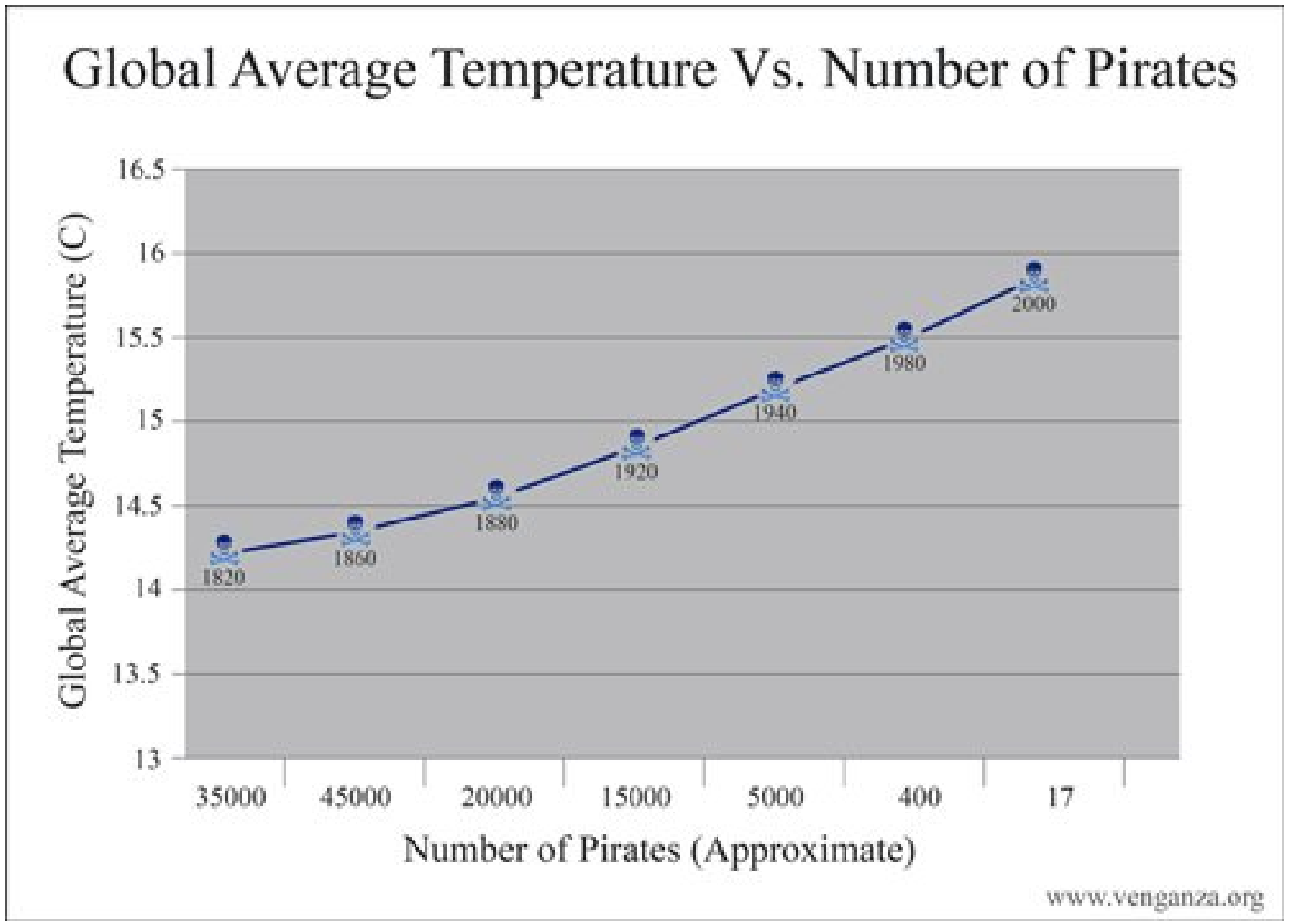
High correlation



A high correlation can be pointed out. Please explain!

Relation between variables

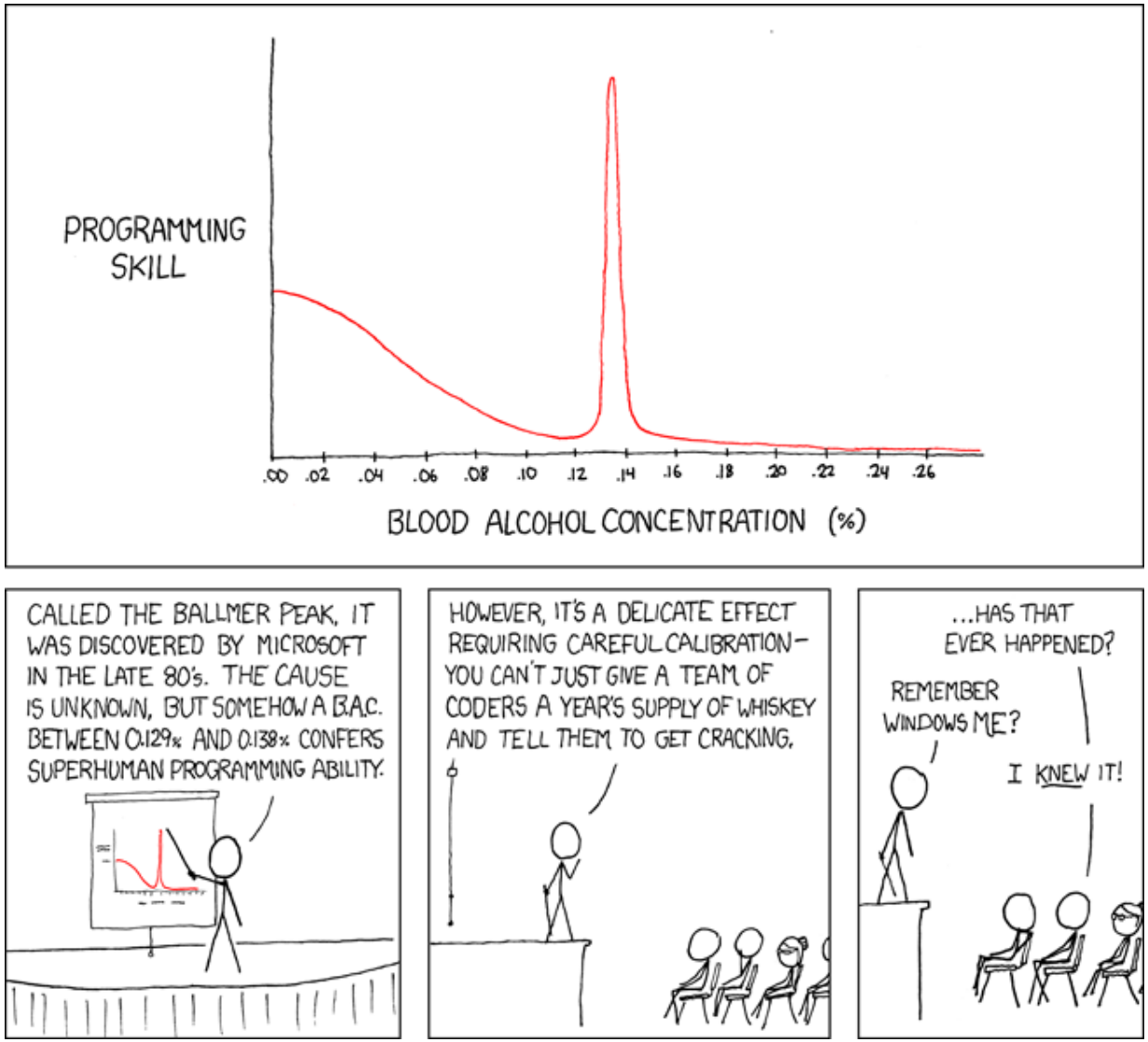
High correlation



A high correlation can be pointed out. So what?

Relation between variables

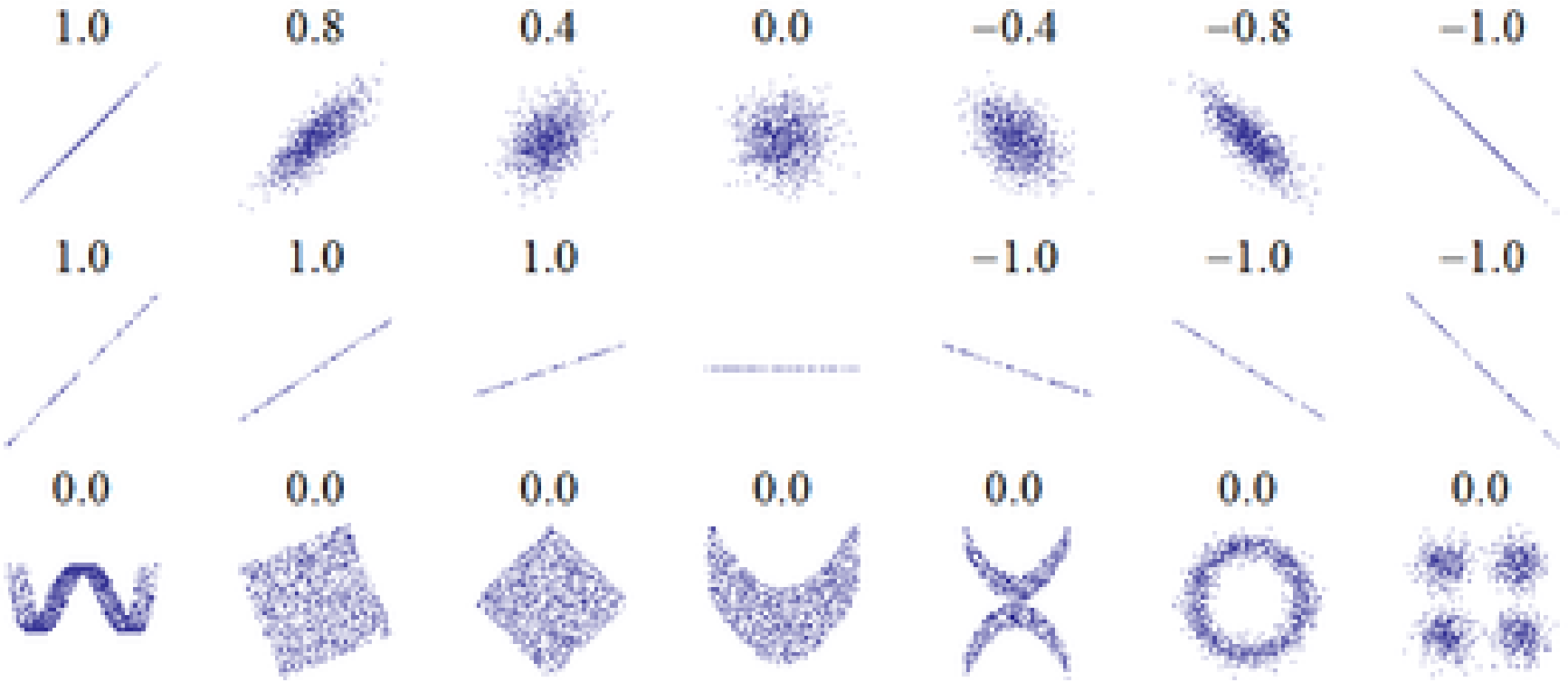
No correlation. No relationship?



Source: *http://xkcd.com/323/*

Relation between variables

Correlation coefficient



Positive (direct: $R = 1$), negative (inverse: $R = -1$), linear, curvilinear and uncorrelated ($R = 0$) relationships

R: correlation coefficient

Research in an elementary school

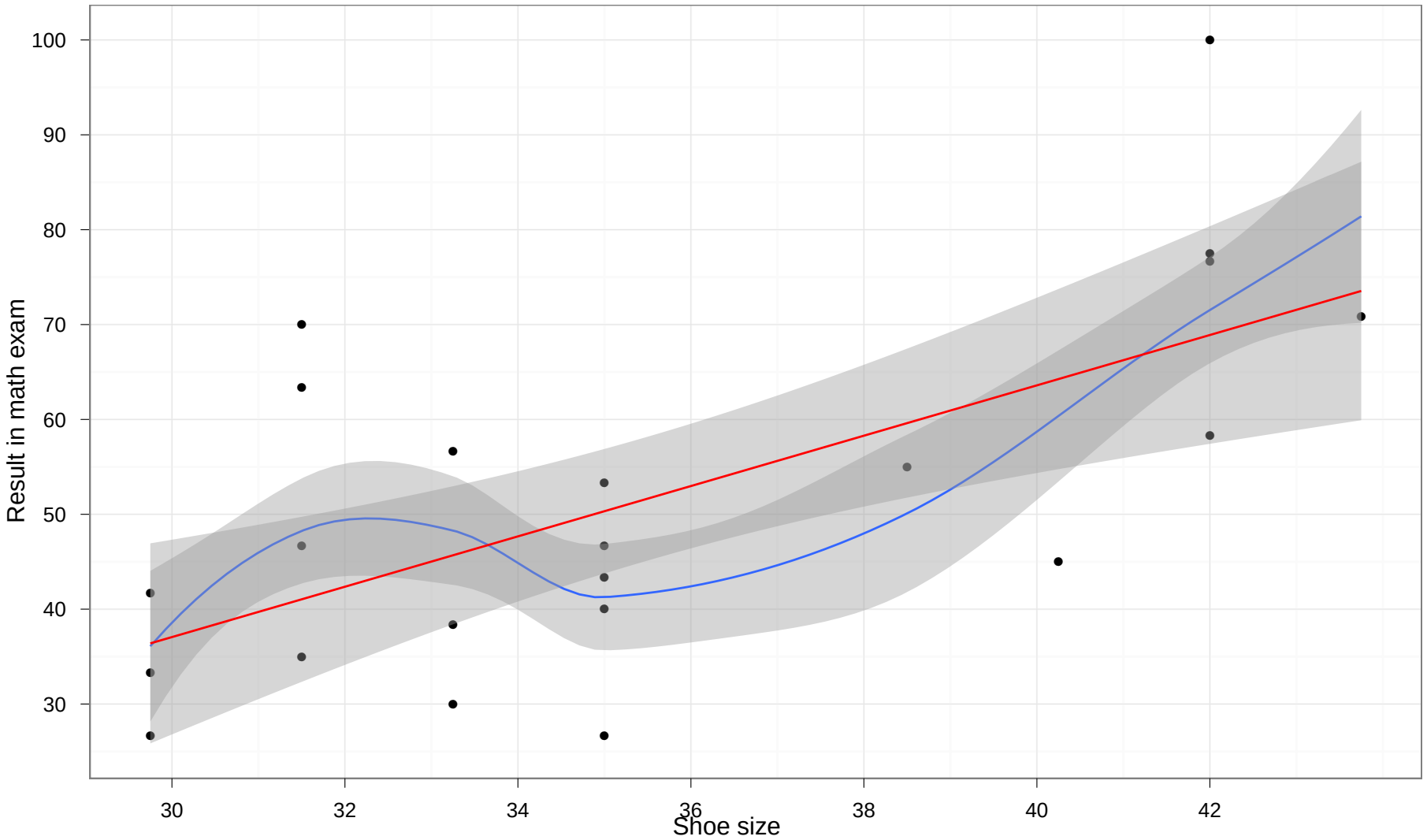
Big shoes and smart kids (example)

We made a small research on the shoe size of some students in an elementary shool, where we also conducted a math exam. See detailed results below:

	Shoe size	Math result
1	29.75	26.67
2	29.75	33.33
3	29.75	41.67
4	31.50	35.00
5	31.50	46.67
6	31.50	63.33
7	31.50	70.00
8	33.25	30.00
9	33.25	38.33
10	33.25	56.67
11	35.00	26.67
12	35.00	40.00
13	35.00	43.33
14	35.00	46.67
15	35.00	53.33
16	38.50	55.00
17	40.25	45.00
18	42.00	58.33
19	42.00	76.67
20	42.00	77.50
21	42.00	100.00
22	43.75	70.83

Research in an elementary school

Big shoes and smart kids (example)



Research in an elementary school

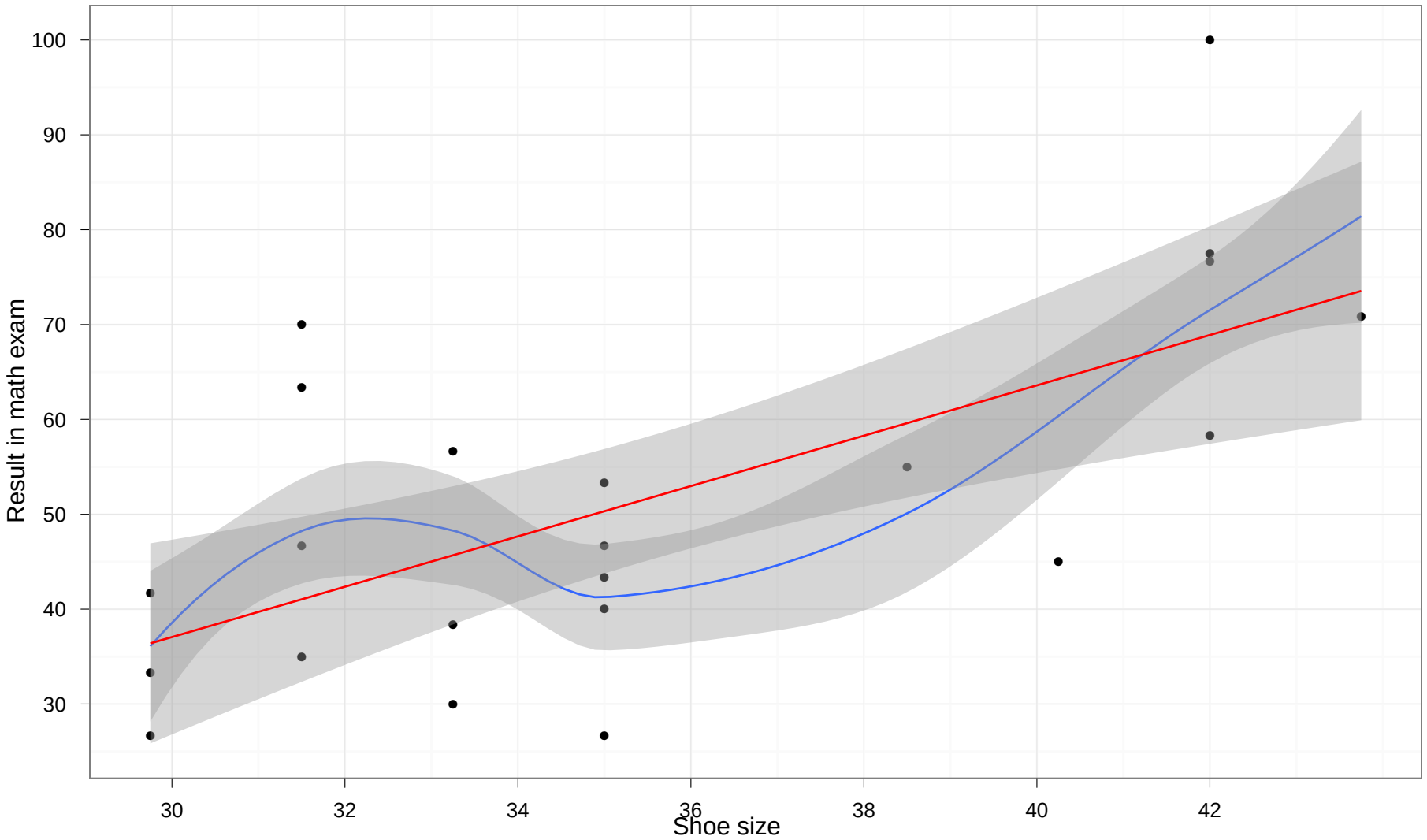
Big shoes and smart kids (example)

We made a small research on the **age** and shoe size of some students in an elementary shool, where we also conducted a math exam. See detailed results below:

	Shoe size	Math result	Age
1	29.75	26.67	3
2	29.75	33.33	7
3	29.75	41.67	5
4	31.50	35.00	8
5	31.50	46.67	10
6	31.50	63.33	11
7	31.50	70.00	12
8	33.25	30.00	7
9	33.25	38.33	7
10	33.25	56.67	12
11	35.00	26.67	6
12	35.00	40.00	8
13	35.00	43.33	6
14	35.00	46.67	10
15	35.00	53.33	11
16	38.50	55.00	9
17	40.25	45.00	9
18	42.00	58.33	9
19	42.00	76.67	16
20	42.00	77.50	18
21	42.00	100.00	19
22	43.75	70.83	14

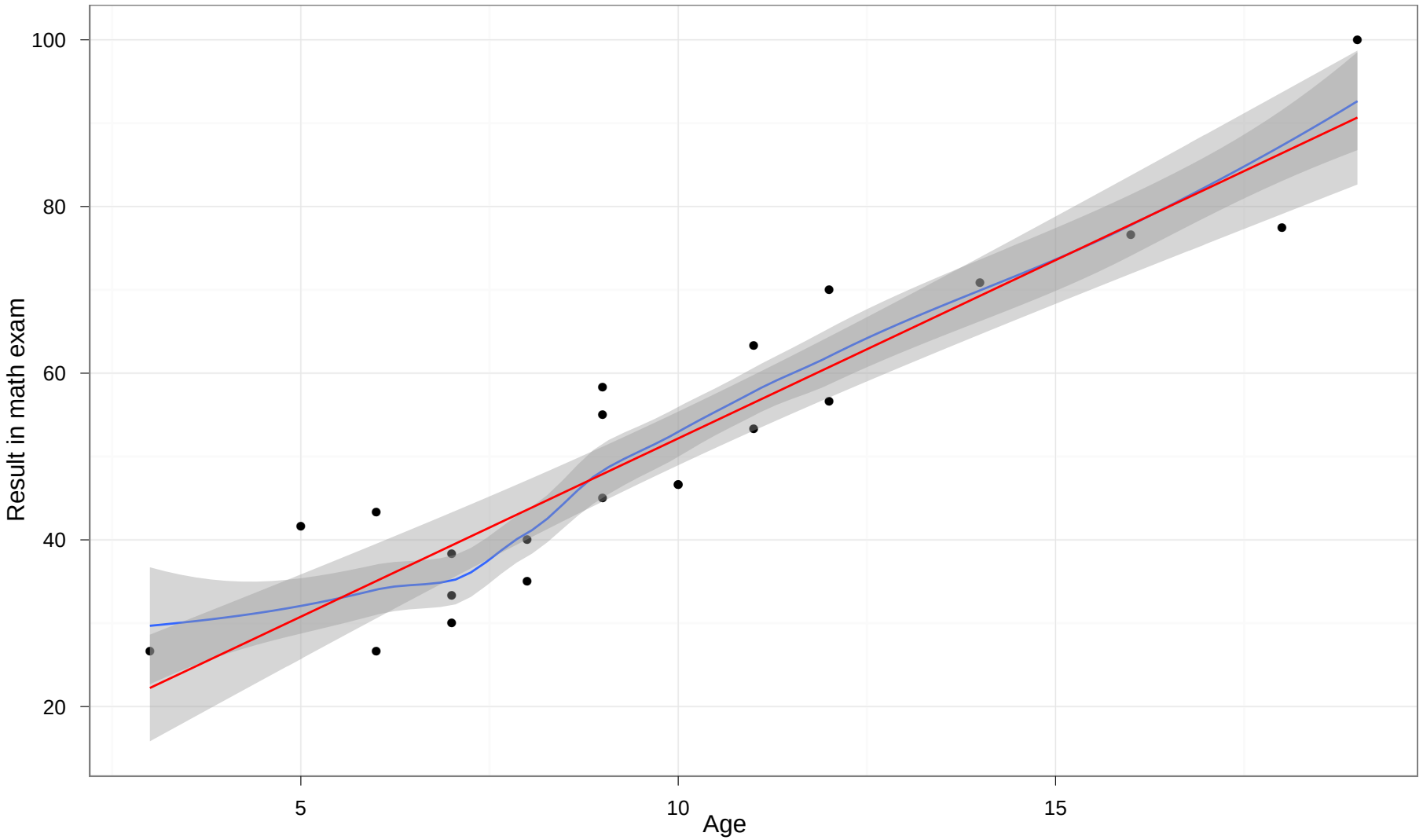
Research in an elementary school

Big shoes and smart kids (example)



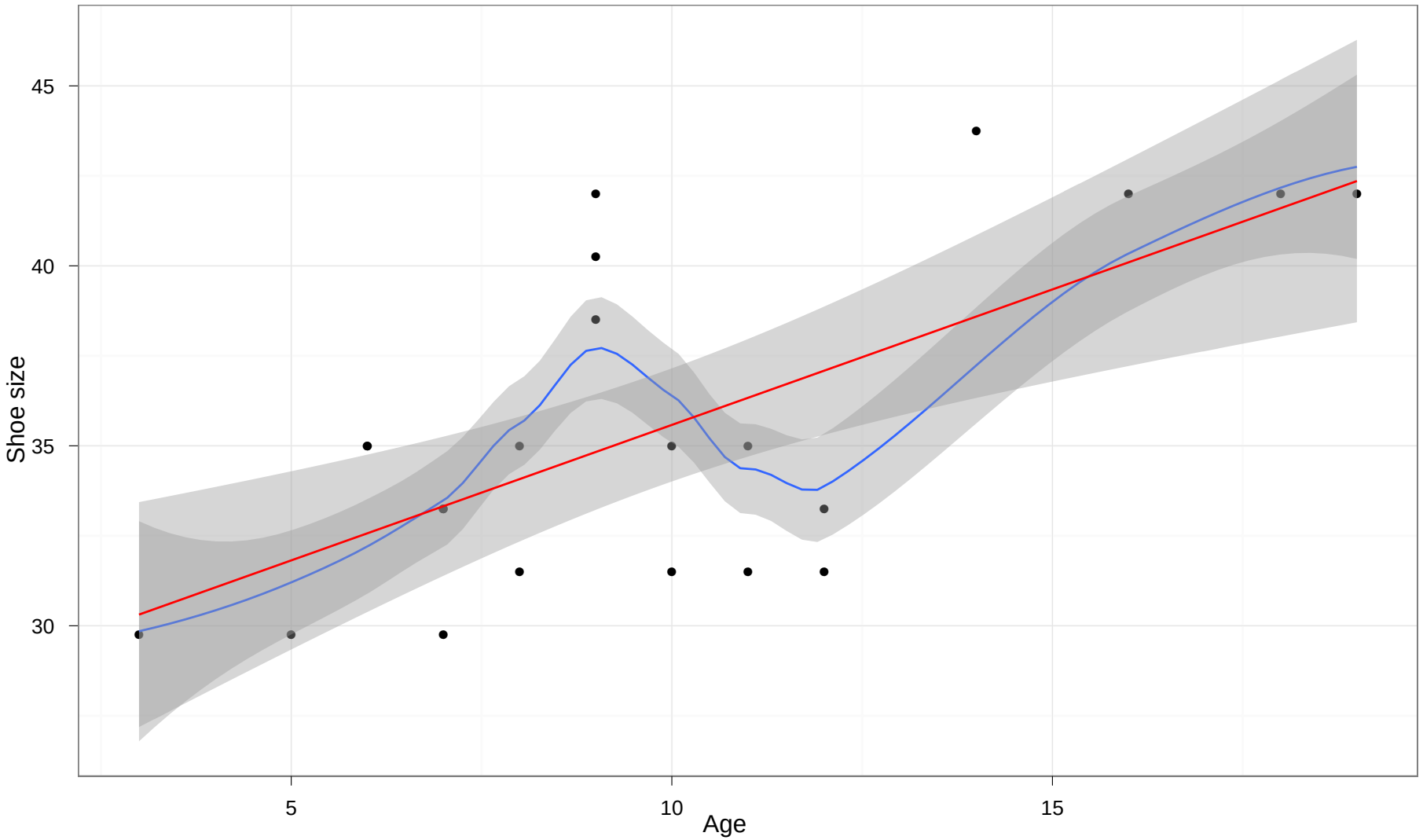
Research in an elementary school

Big shoes and smart kids (example)



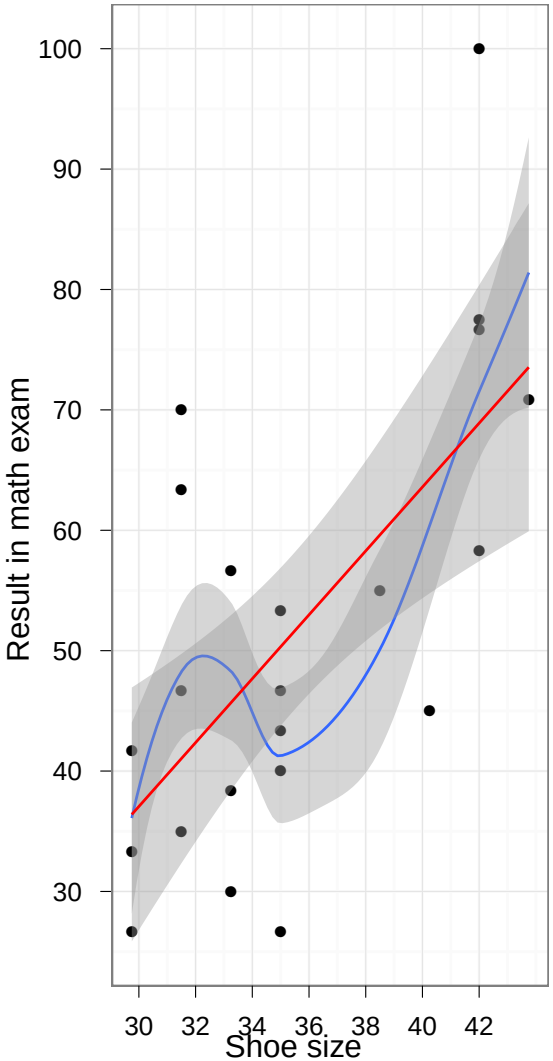
Research in an elementary school

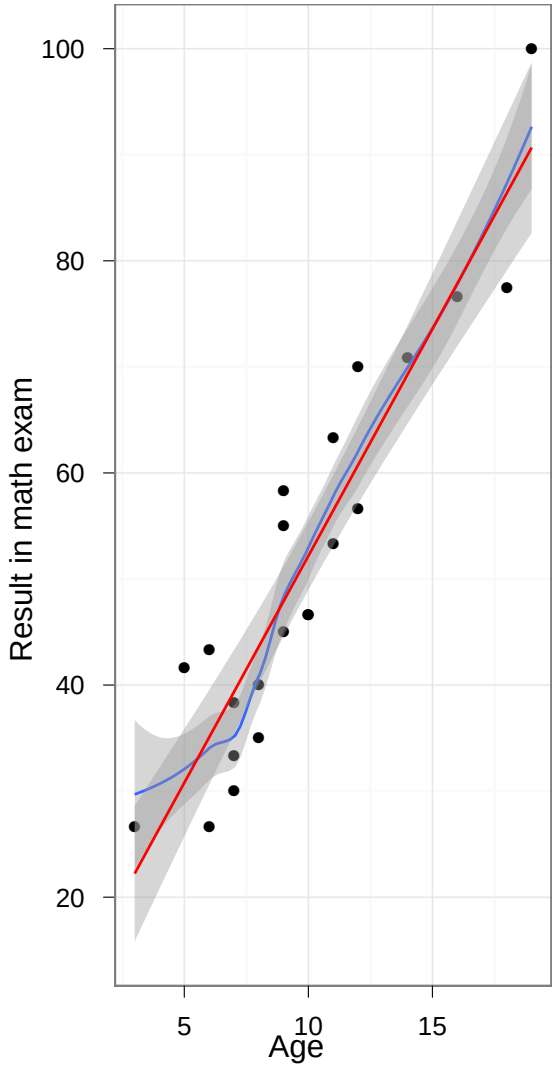
Big shoes and smart kids (example)

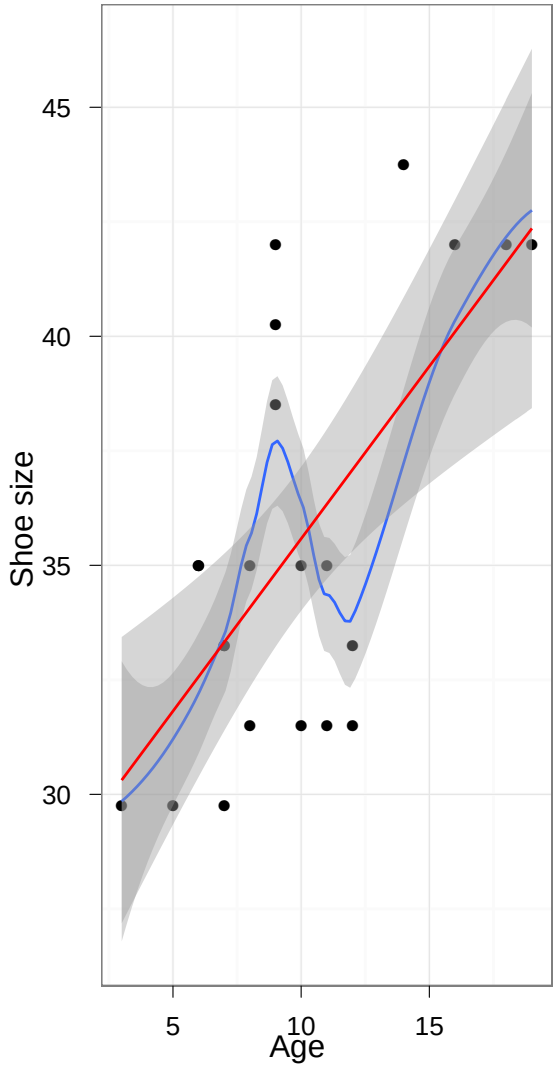


Research in an elementary school

Big shoes and smart kids (example)

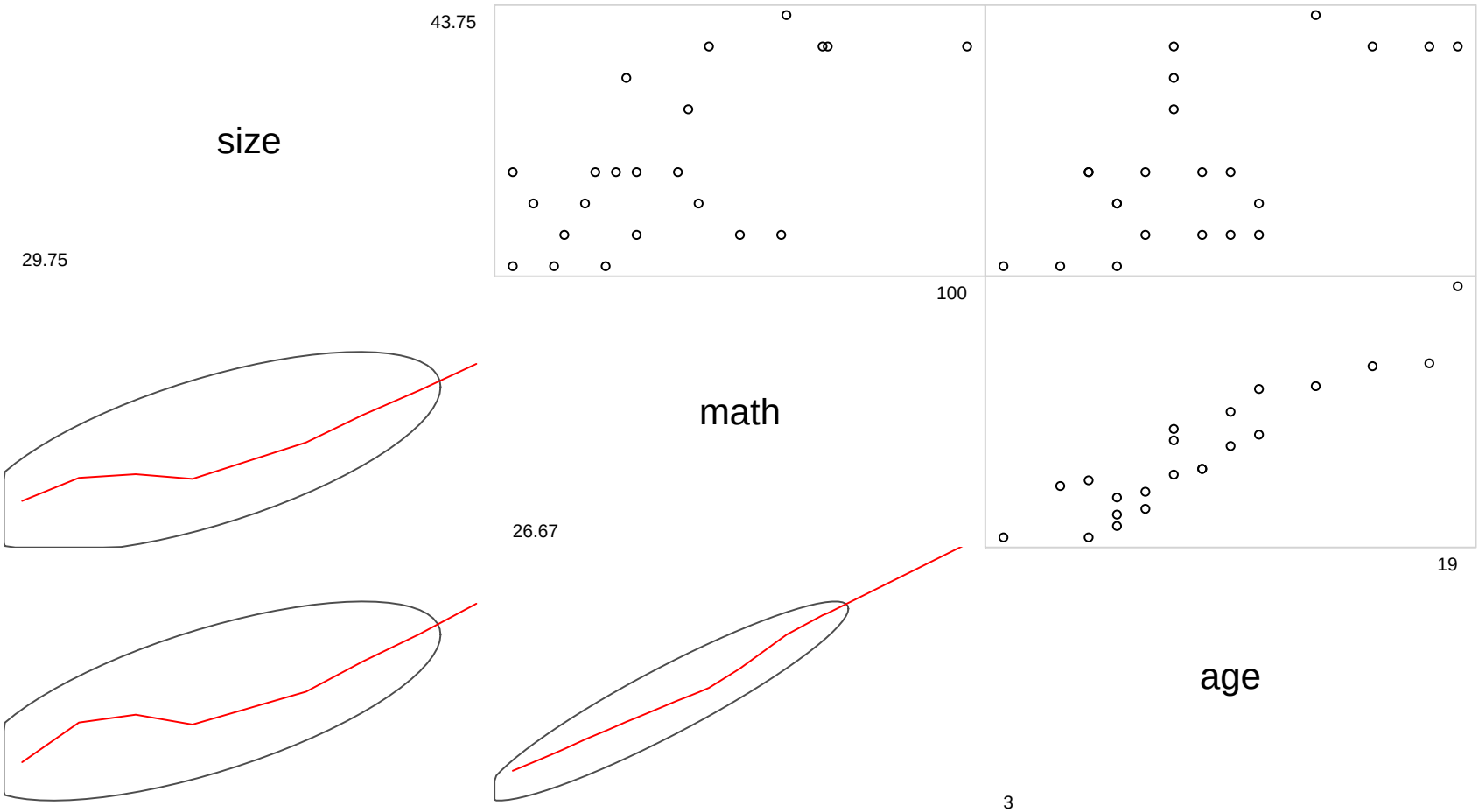






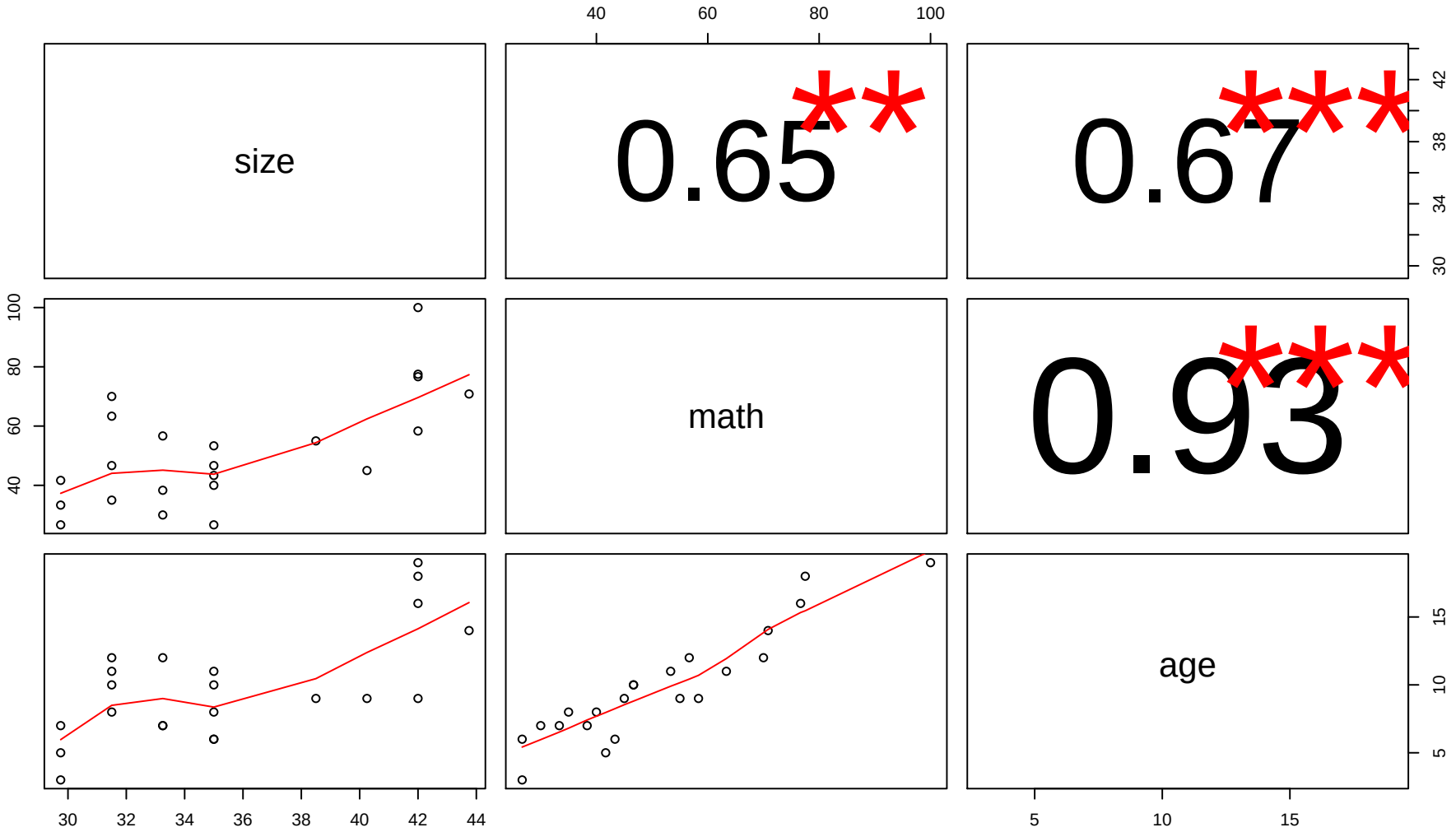
Research in an elementary school

Big shoes and smart kids (example)



Research in an elementary school

Big shoes and smart kids (example)



Partial correlation:

$$r_{math,size \cdot age} = 0.11$$

$$r_{math,age \cdot size} = 0.87$$

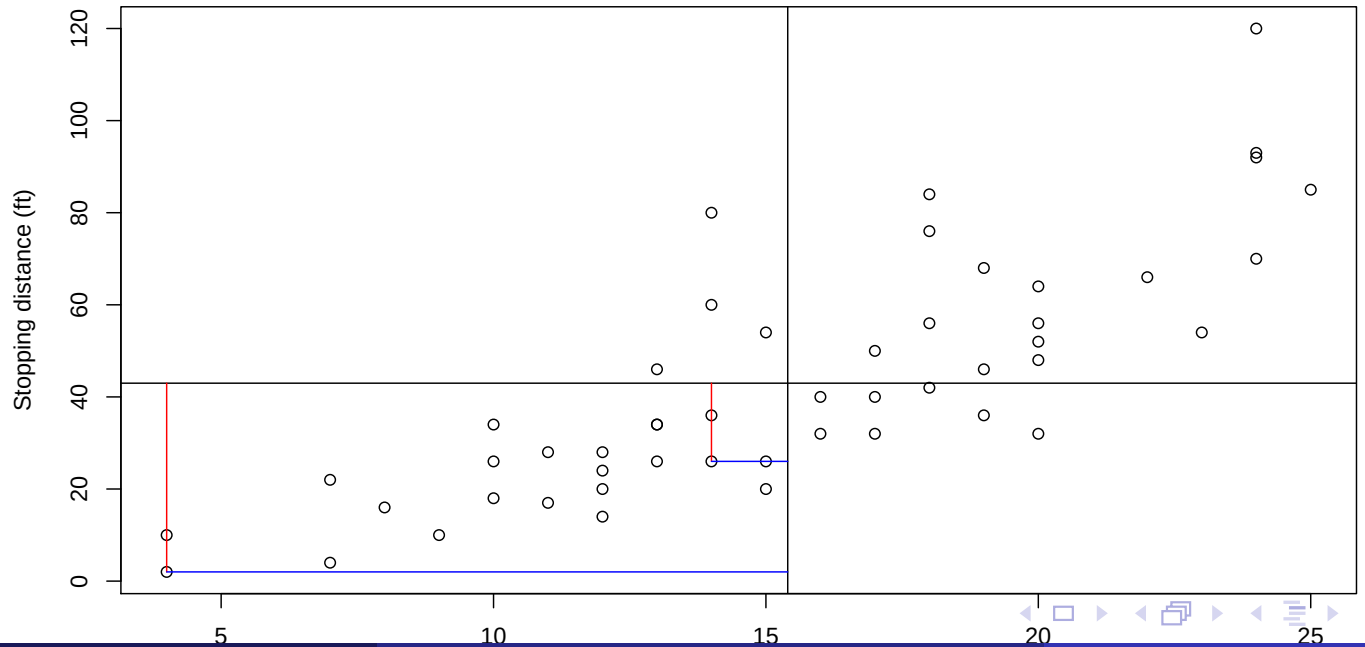
$$r_{size,age \cdot math} = 0.22$$

For x and y variables the joint variablity could be computed by :

$$COV(xy) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$remember : \sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$$

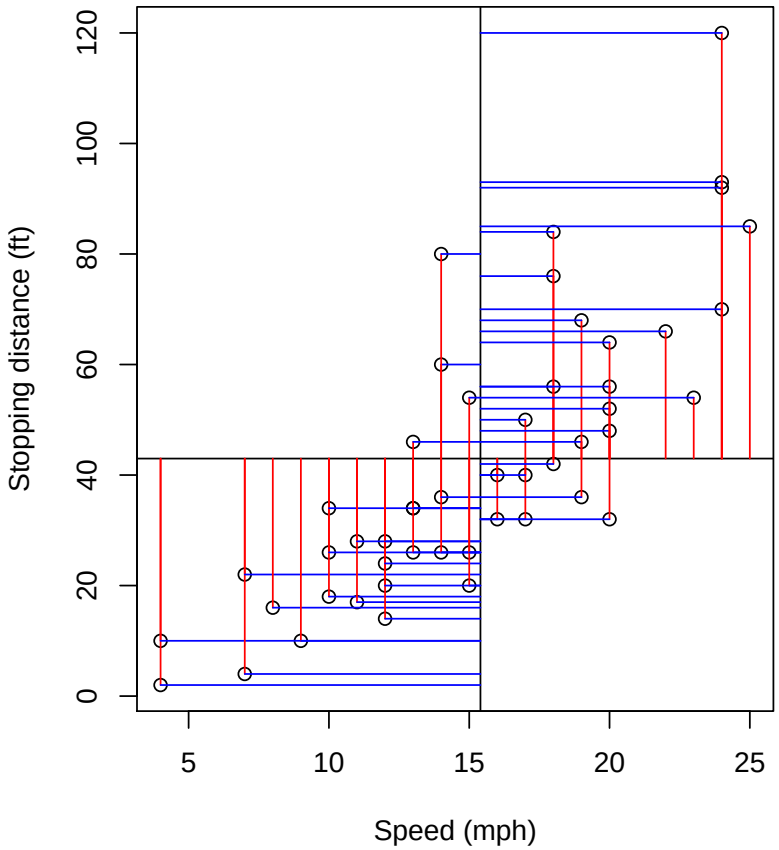
Ezekiel, M. (1930) Methods of Correlation Analysis. Wiley.



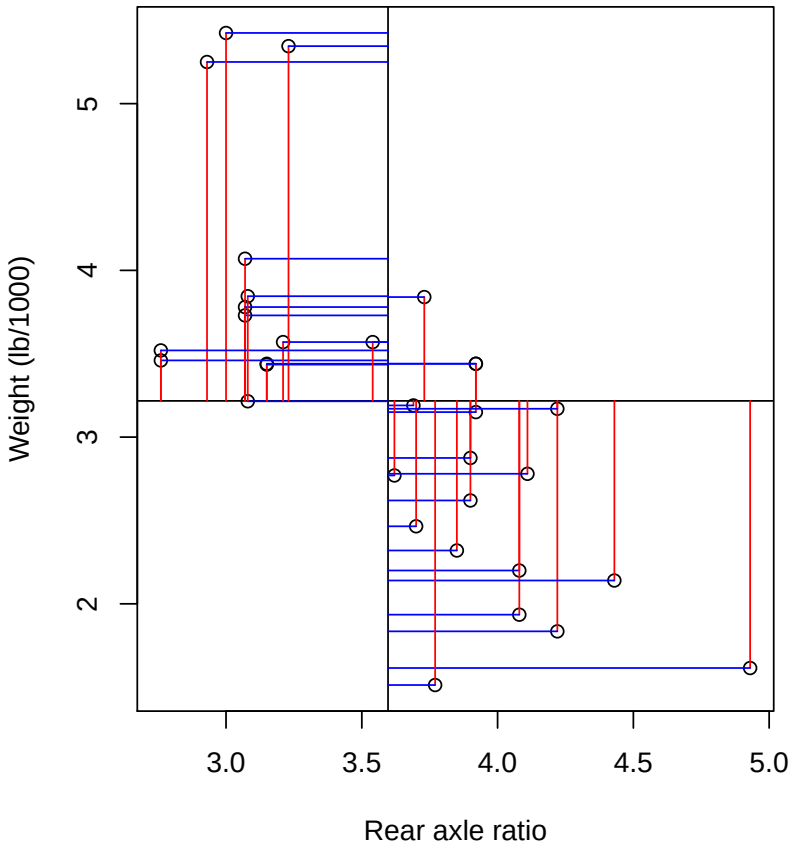
Theoretical background

Covariation

Ezekiel, M. (1930):
Methods of Correlation Analysis



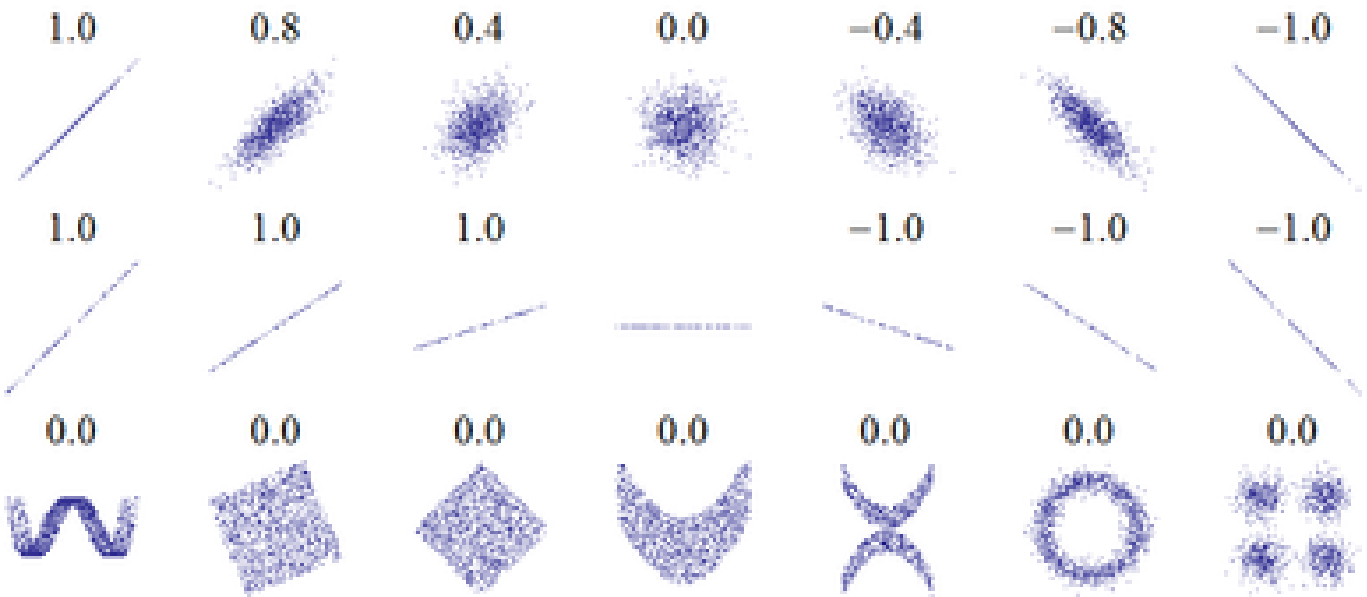
Henderson & Velleman (1981):
Building multiple regression models interactively



Theoretical background

Correlation

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



$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{N \sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \sum_{i=1}^{N} r_{X,i}^2 - \left(\sum_{i=1}^{N} r_{X,i} \right)^2} \sqrt{N \sum_{i=1}^{N} r_{Y,i}^2 - \left(\sum_{i=1}^{N} r_{Y,i} \right)^2}}$$

so for three variables:

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{r_{XY} - r_{X\mathbf{Z}} r_{Y\mathbf{Z}}}{\sqrt{(1 - r_{X\mathbf{Z}}^2)(1 - r_{Y\mathbf{Z}}^2)}}$$

Exercises

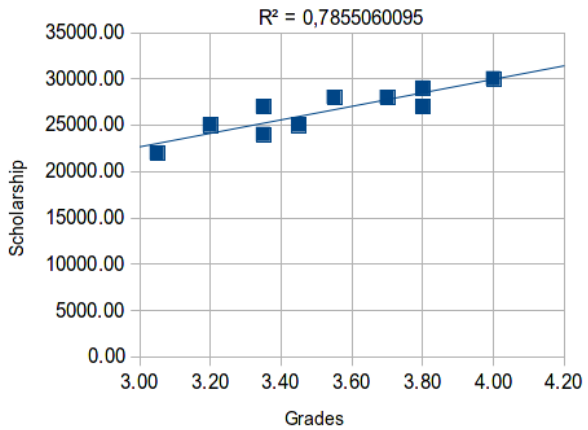
1. What is correlation and partial correlation?
2. Building upon your findings, compute the possible pairs of correlation coefficients on the below dataset!
3. Also look for partial correlation and comment on your results!

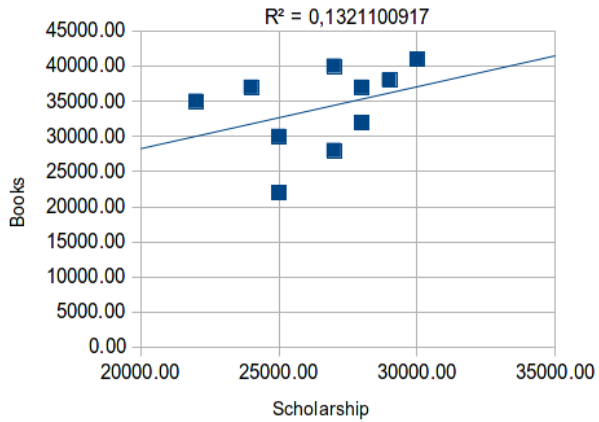
Grade (mean)	Scholarship (in HUF)	Money spent on books (in HUF)
3.05	22000	3500
3.2	25000	3000
3.35	27000	2800
3.35	24000	3700
3.45	25000	2200
3.55	28000	3200
3.7	28000	3700
45	30000	4100
3.8	27000	4000
3.8	29000	3800

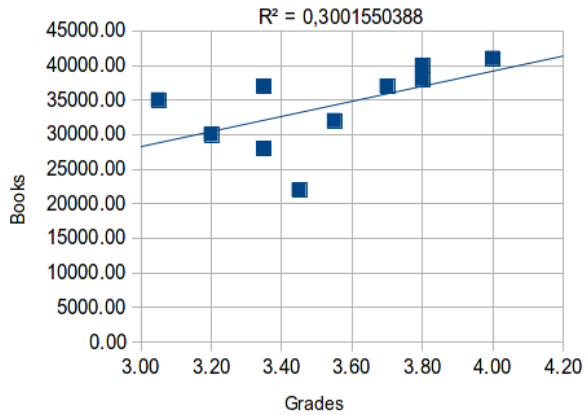
Exercises

Solution

				$x_i - \bar{x}$			$(x_i - \bar{x})^2$			$(x_i - \bar{x})(y_i - \bar{y})$		
	Grade	Scholarship	Books	Grade	Scholarship	Books	Grade	Scholarship	Books	Grade-sch	Sch-books	Grade-books
	3.05	22000.00	35000.00	-0.48	-4500.00	1000.00	0.225625	20250000	1000000	2137.5	-4500000	-475
	3.20	25000.00	30000.00	-0.33	-1500.00	-4000.00	0.105625	2250000	16000000	487.5	6000000	1300
	3.35	27000.00	28000.00	-0.18	500.00	-6000.00	0.030625	250000	36000000	-87.5	-3000000	1050
	3.35	24000.00	37000.00	-0.18	-2500.00	3000.00	0.030625	6250000	9000000	437.5	-7500000	-525
	3.45	25000.00	22000.00	-0.07	-1500.00	-12000.00	0.005625	2250000	144000000	112.5	18000000	900
	3.55	28000.00	32000.00	0.02	1500.00	-2000.00	0.000625	2250000	4000000	37.5	-3000000	-50
	3.70	28000.00	37000.00	0.18	1500.00	3000.00	0.030625	2250000	9000000	262.5	4500000	525
	4.00	30000.00	41000.00	0.48	3500.00	7000.00	0.225625	12250000	49000000	1662.5	24500000	3325
	3.80	27000.00	40000.00	0.28	500.00	6000.00	0.075625	250000	36000000	137.5	3000000	1650
	3.80	29000.00	38000.00	0.28	2500.00	4000.00	0.075625	6250000	16000000	687.5	10000000	1100
Min	3.05	22000.00	22000.00			Sum	0.80625	54500000	320000000	5875	48000000	8800
Max	4.00	30000.00	41000.00			St. dev.	0.29930475	2460.80384	5962.84794	r	0.89	0.36	0.55
Range	0.95	8000.00	19000.00							r^2	0.79	0.13	0.30
Mean	3.53	26500.00	34000.00							partial corr	0.96	-0.24	0.46
Median	3.50	27000.00	36000.00										





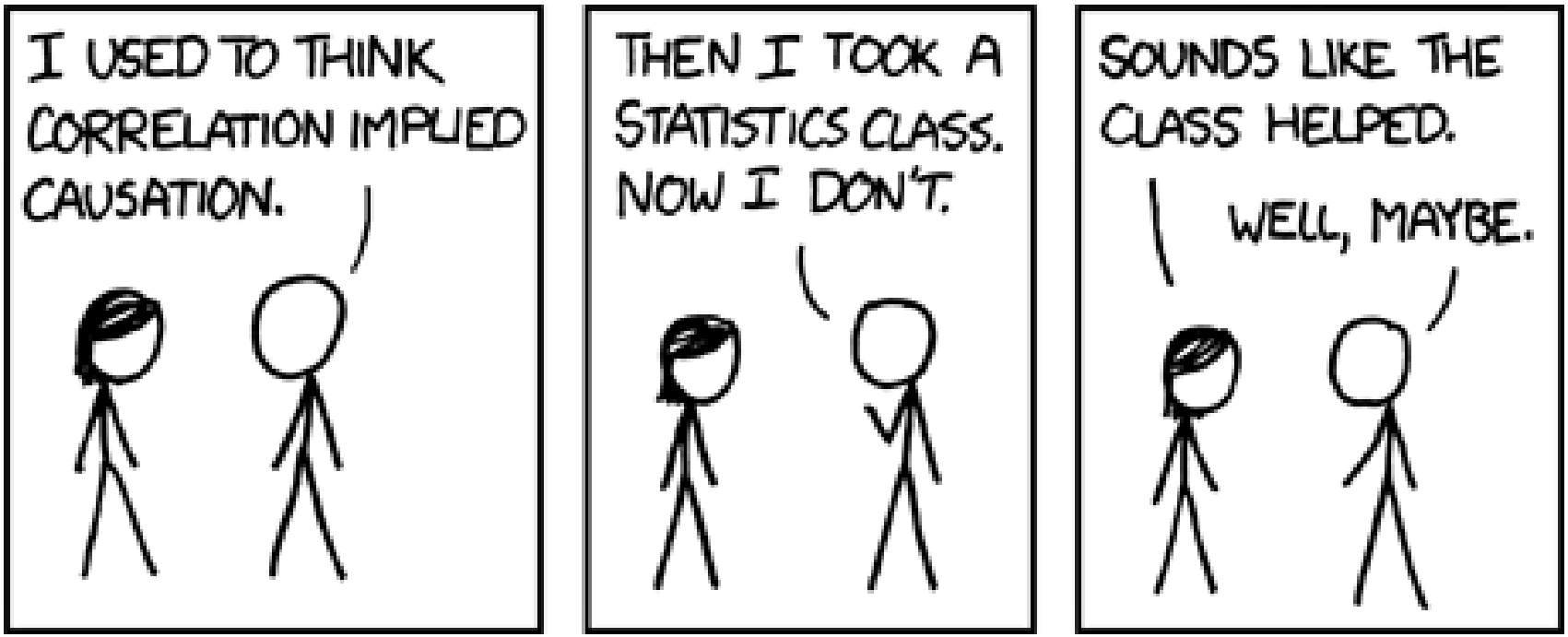


Limitations of the correlation coefficient

- Correlation and causality
- Lazarsfeld paradigm
- Correlation and linearity

Limitations of the correlation coefficient

Correlation does not imply causation!



Source: http://xkcd.com/552

Limitations of the correlation coefficient

Correlation does not imply causation! - Theoretical background

Aristotle: logic, syllogism – if $(A \rightarrow B)\&(B \rightarrow C) \Rightarrow A \rightarrow C$

David Hume: scepticism

- "only correlation can actually be perceived [not causality]"
- see: our belief that the sun will rise tomorrow
- see: "If I see a billiard ball moving towards another, on a smooth table, I can easily conceive to stop upon contact."

Popper: falsification

Pearl, J. - *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000

Stouffer: *The American Soldier*

Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion.

Stouffer: *The American Soldier*

H_0: Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. **BUT:**

"Soldiers in branches with higher promotion rates were more pessimistic about their own chances of being promoted than soldiers in branches with lower rates of promotion."

Limitations of the correlation coefficient

Lazarsfeld paradigm

Stouffer: *The American Soldier*

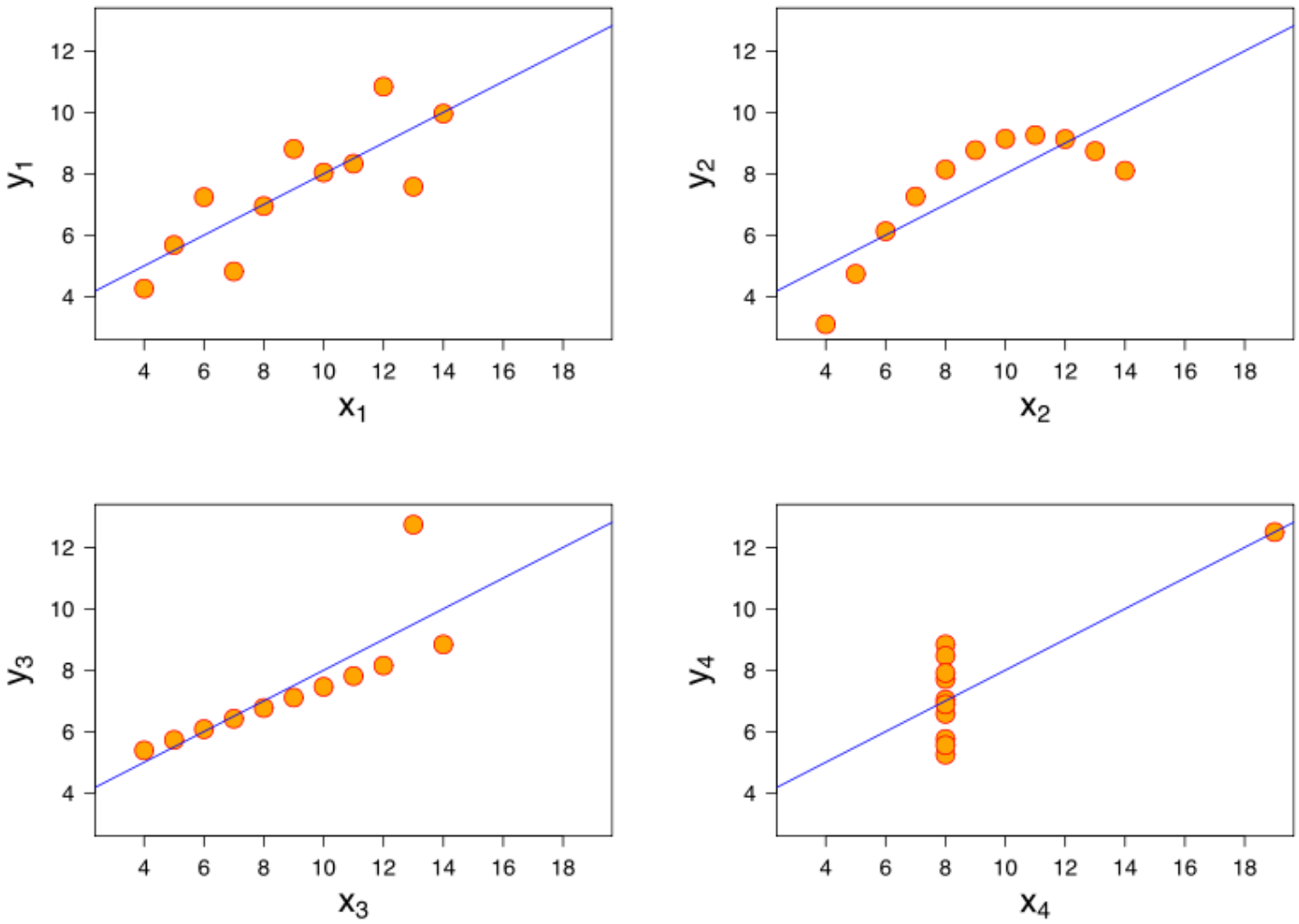
H_0: Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. **BUT:**

"Soldiers in branches with higher promotion rates were more pessimistic about their own chances of being promoted than soldiers in branches with lower rates of promotion."

Keywords: **reference group**, **relative deprivation**

Limitations of the correlation coefficient

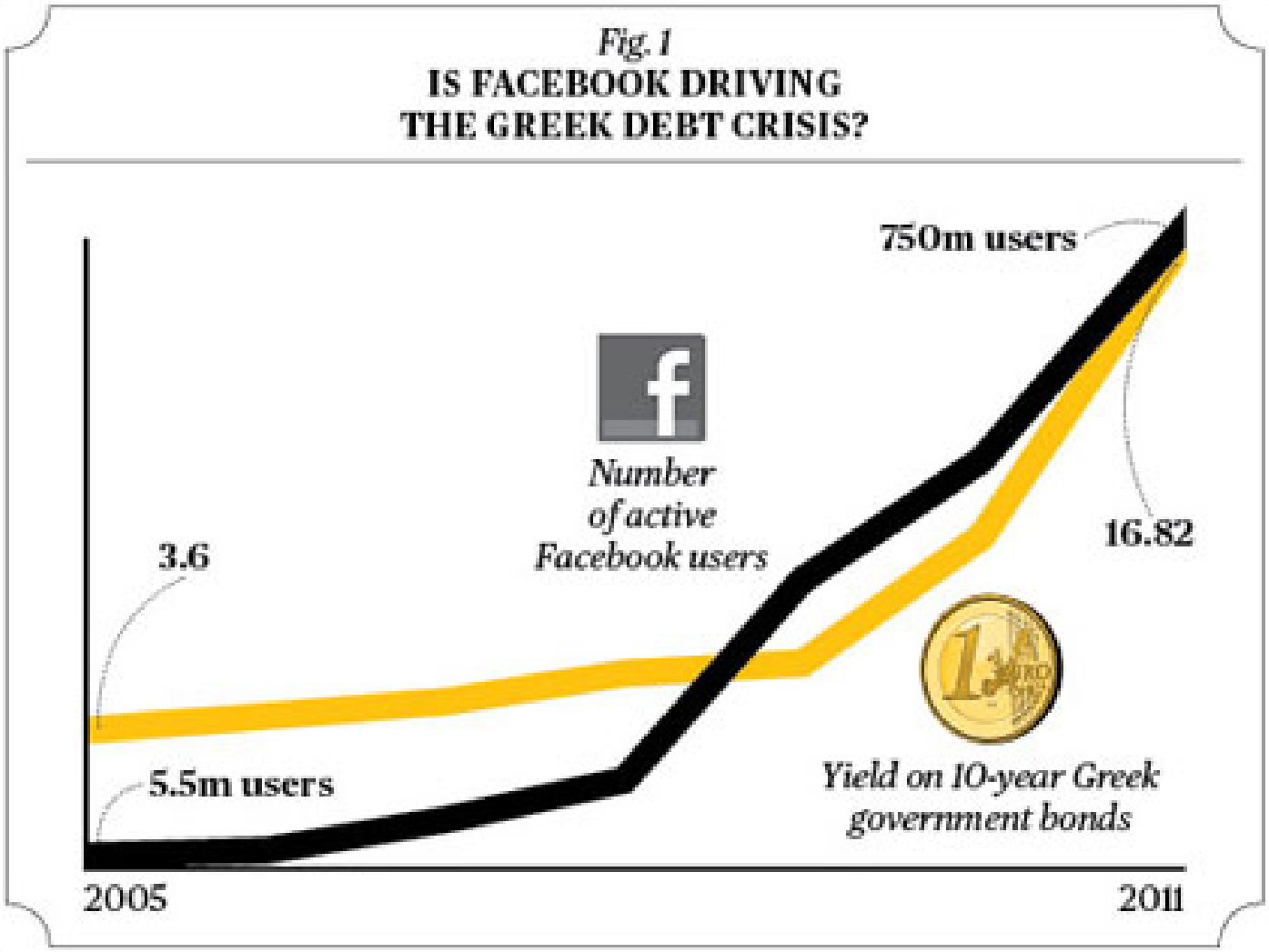
Correlation and linearity - Variations of the Same Theme



Source: Anscombe, F. J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

Relation between variables

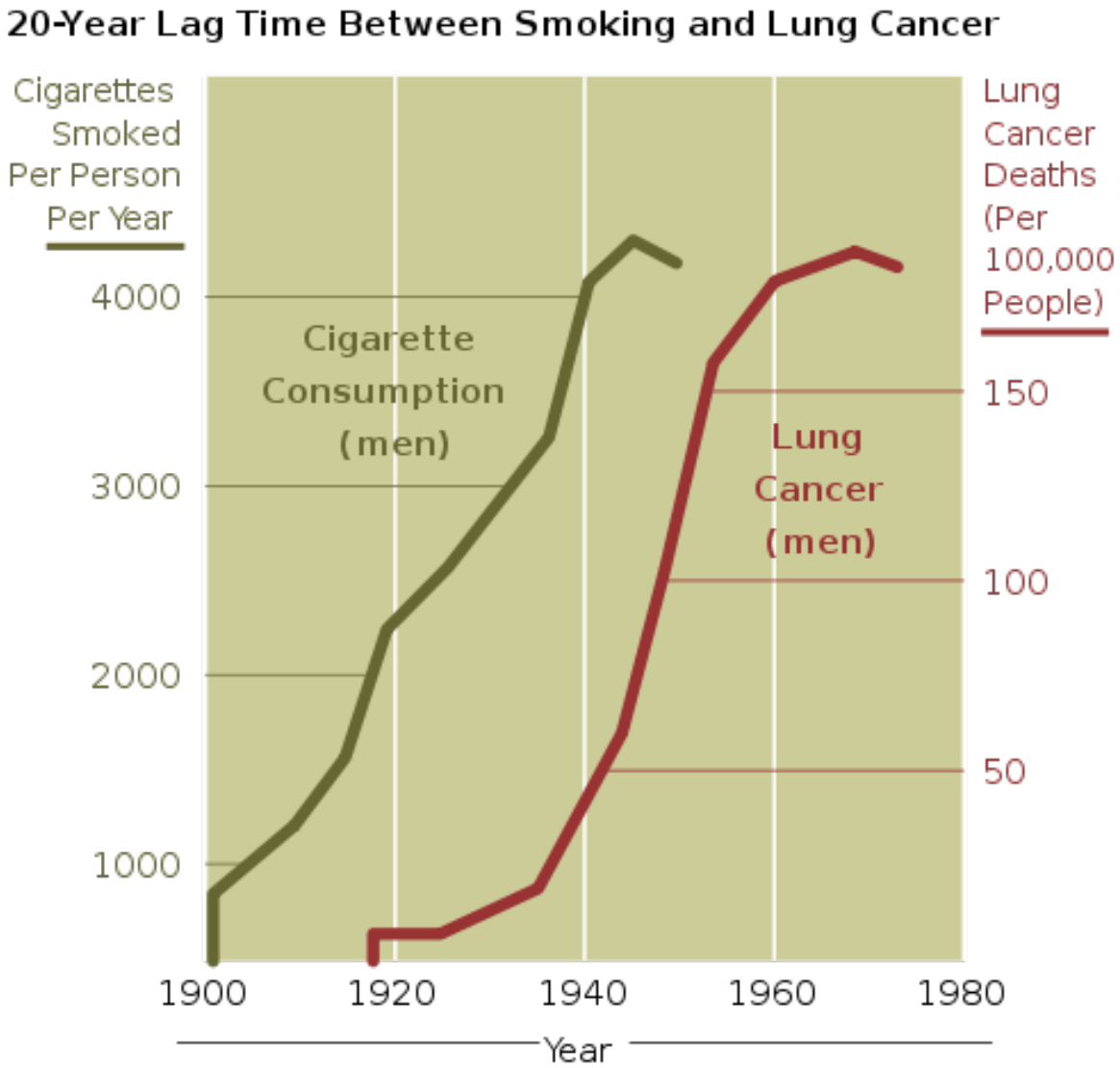
Correlation



Source: http://www.businessweek.com/magazine/correlation-or-causation-12012011-gfx.html

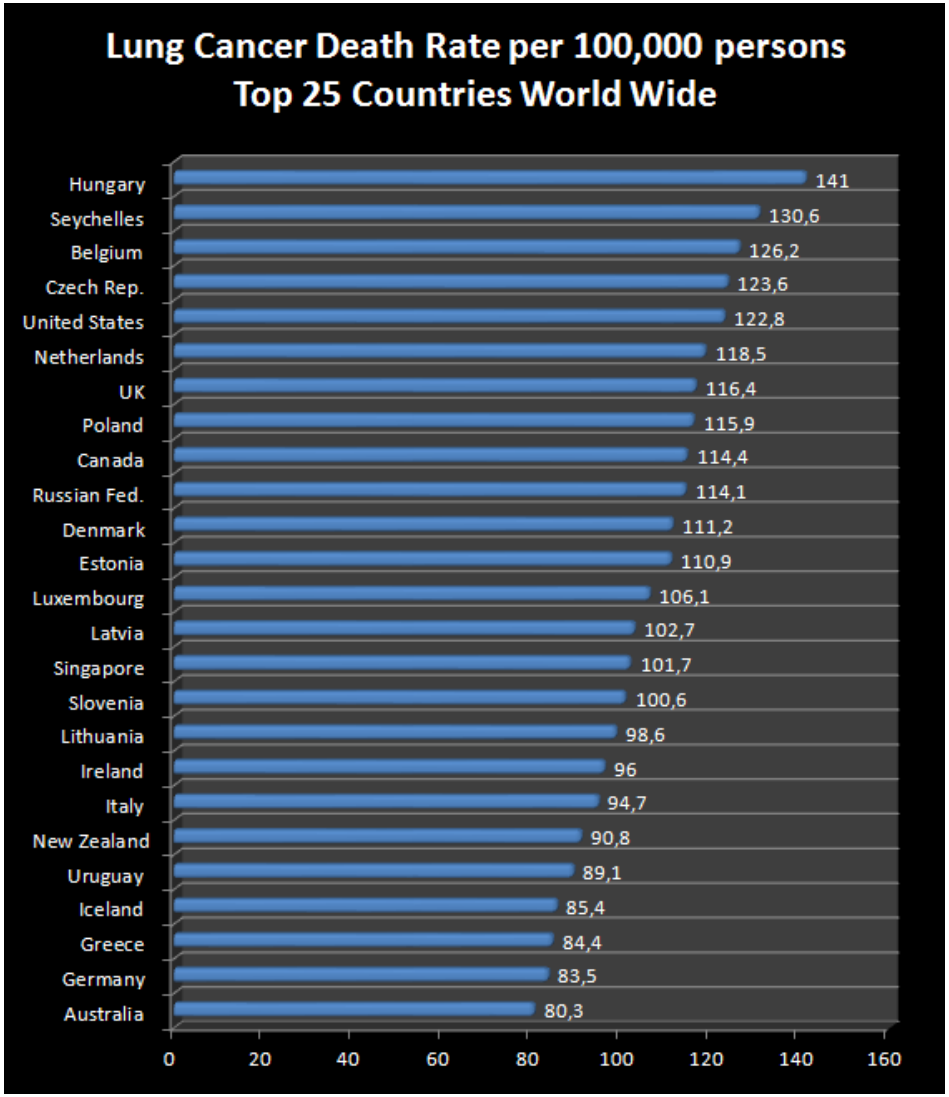
Relation between variables

Cross-correlation



Relation between variables

?



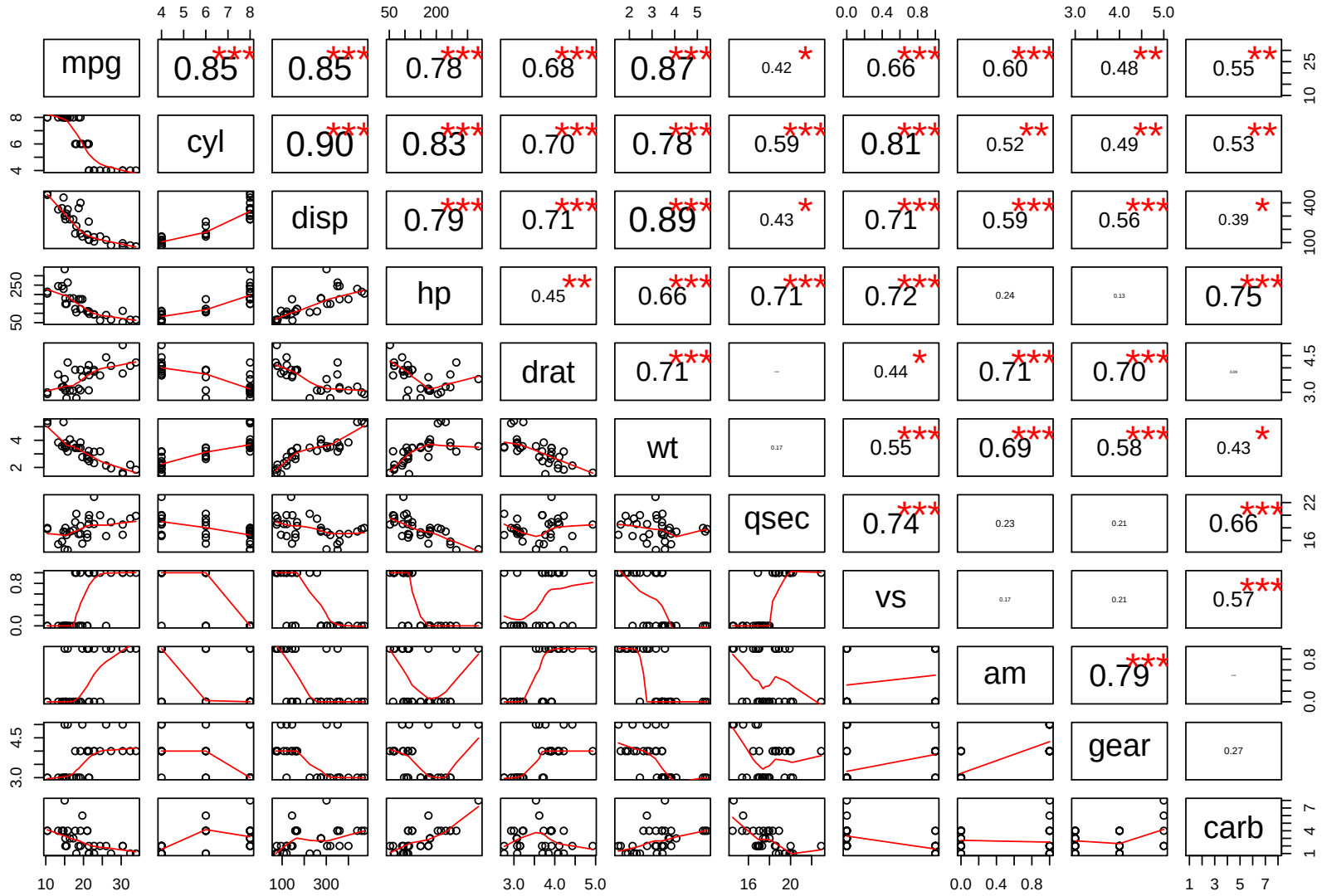
Exercise

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (lb/1000)
- qsec: 1/4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

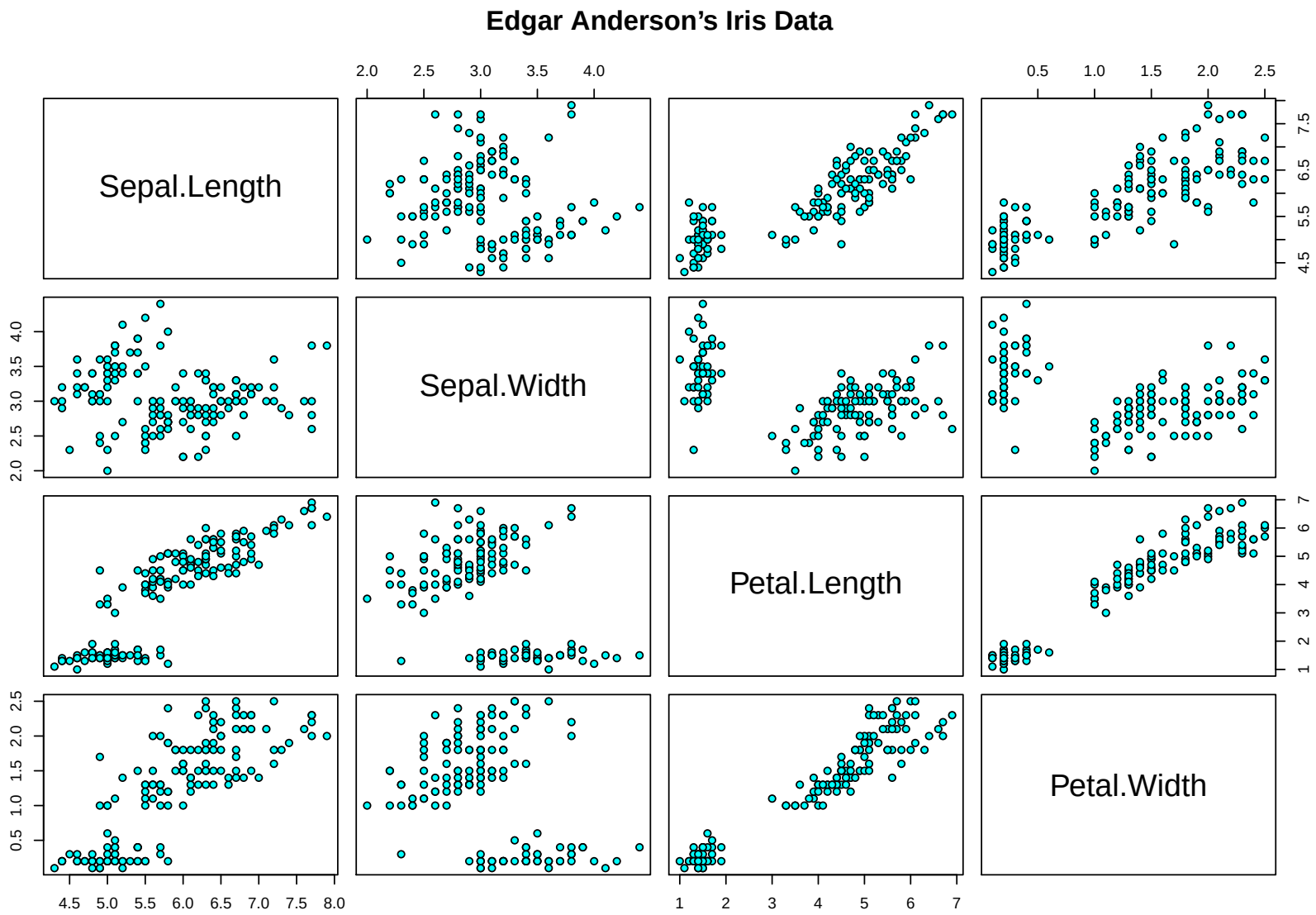
Source: Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391-411.

Exercise



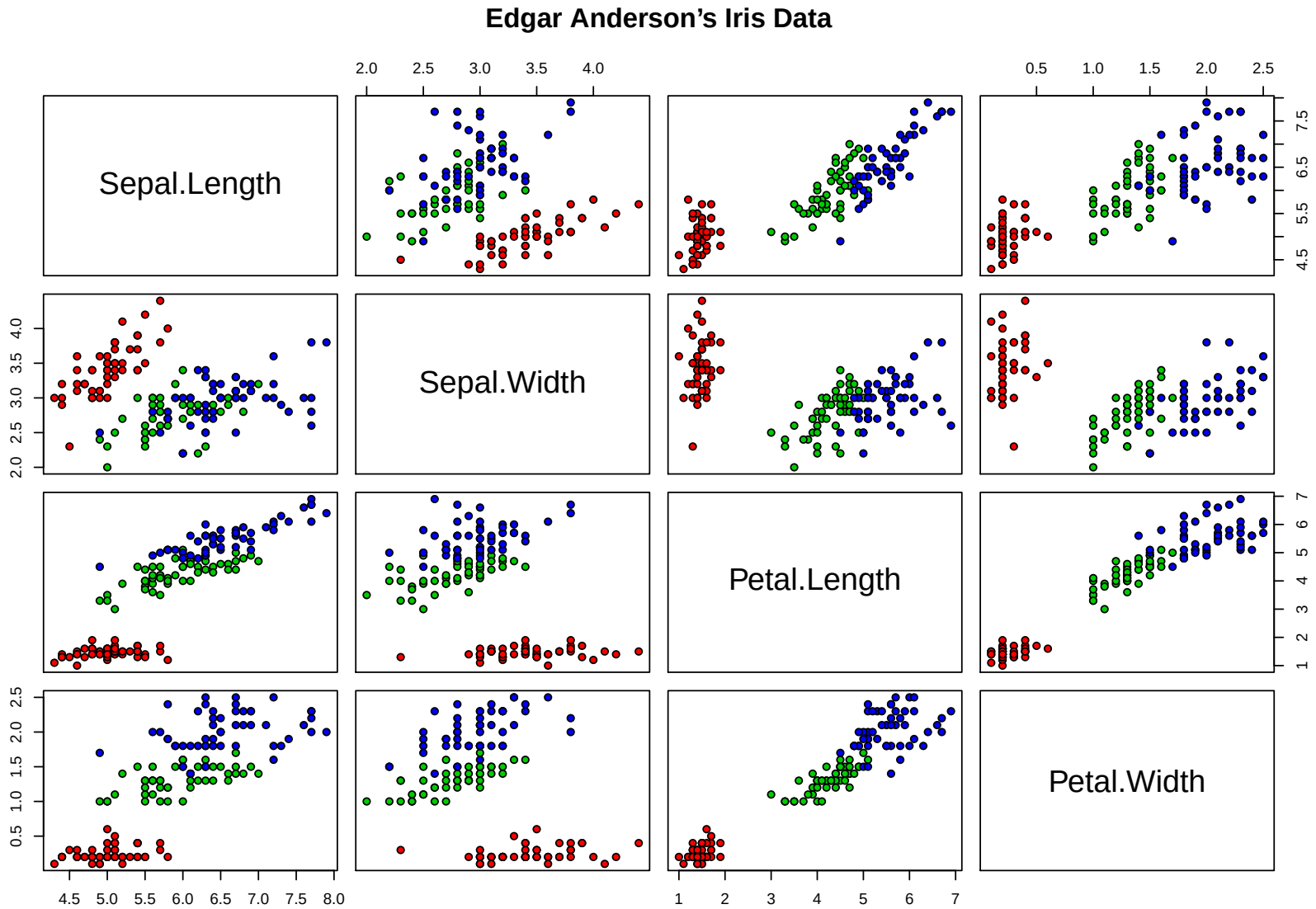
Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391-411.

Exercise



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, 59, 2-5.

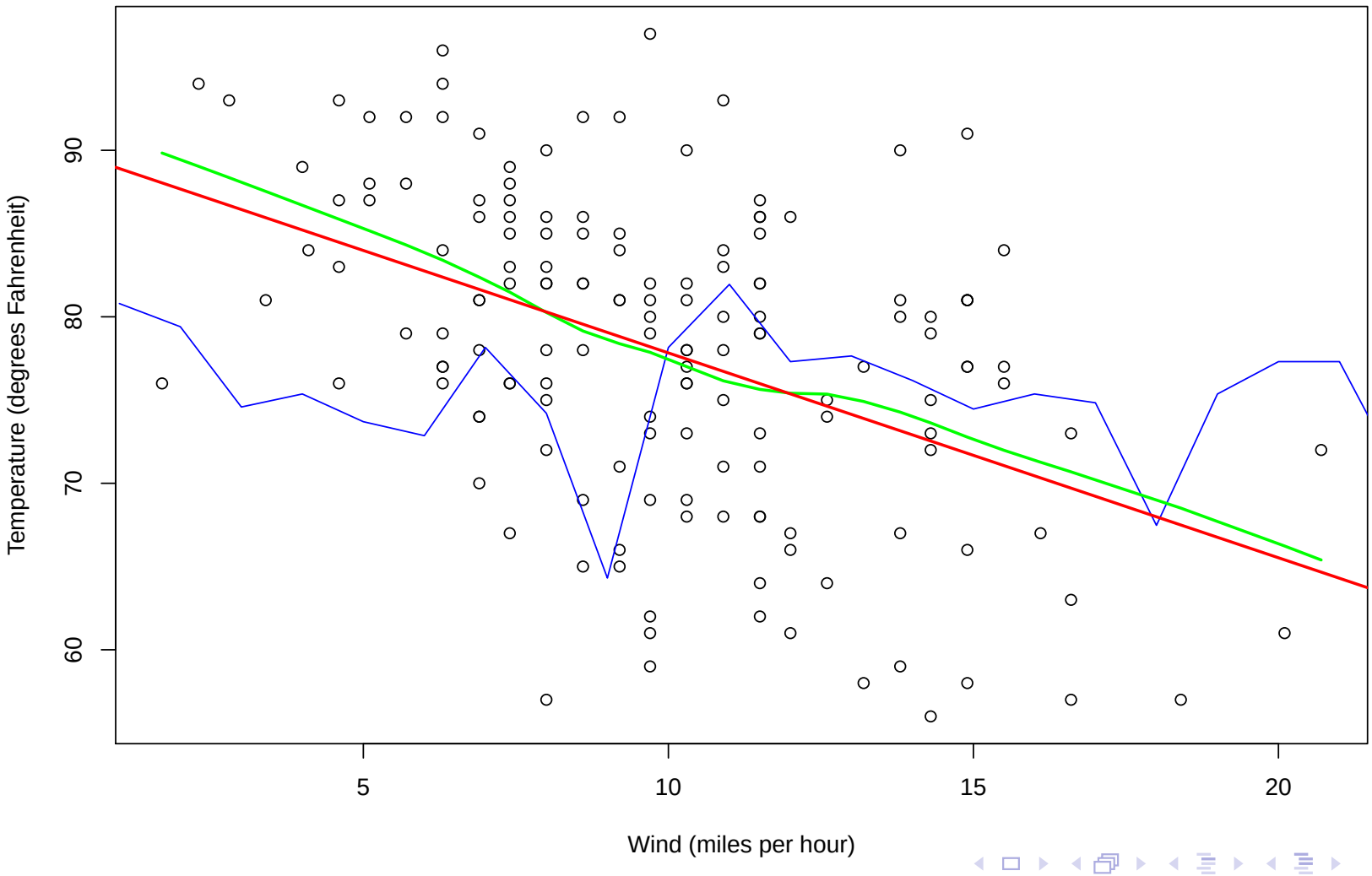
Exercise #2



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, 59, 2-5.

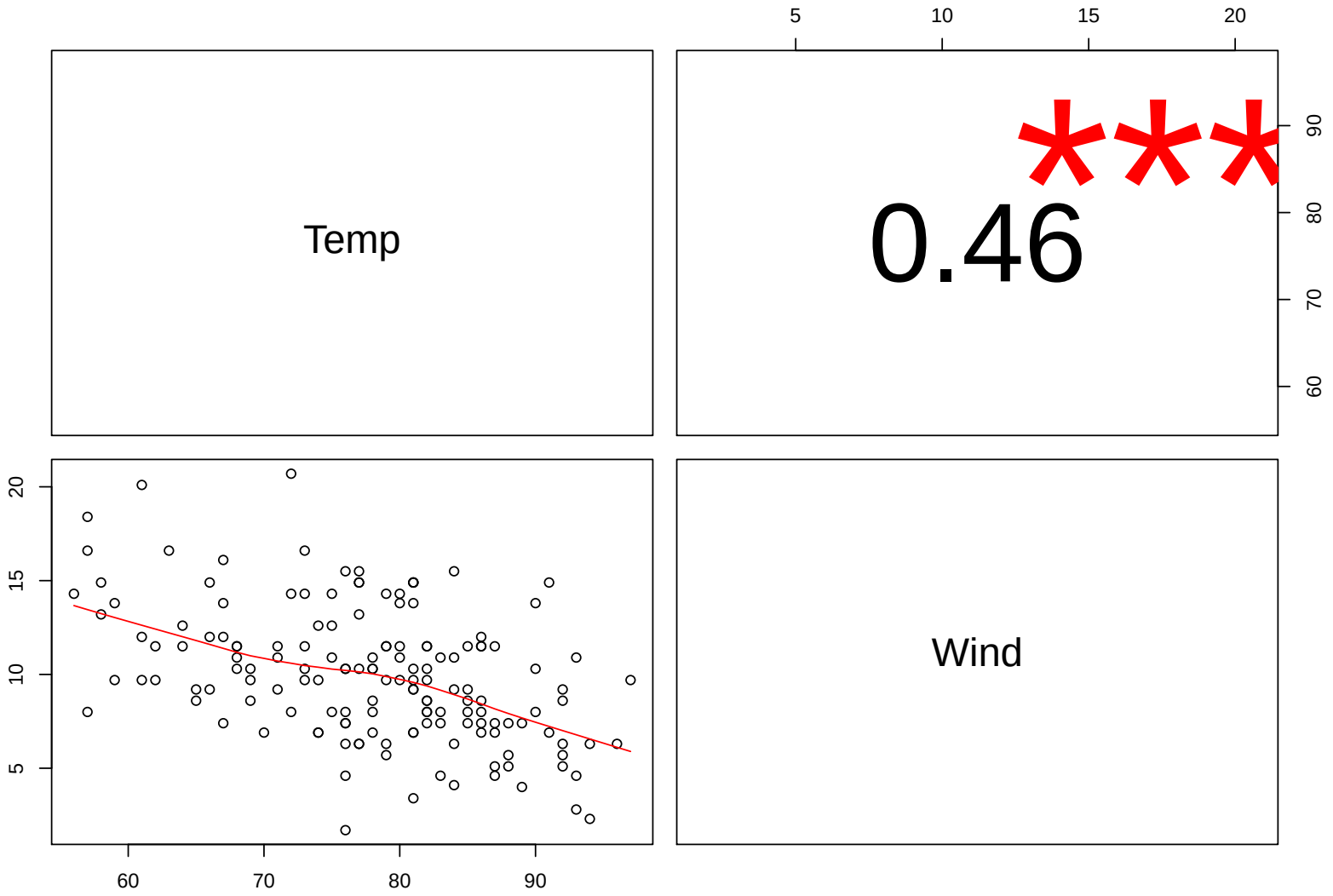
Exercise #3

Real association?



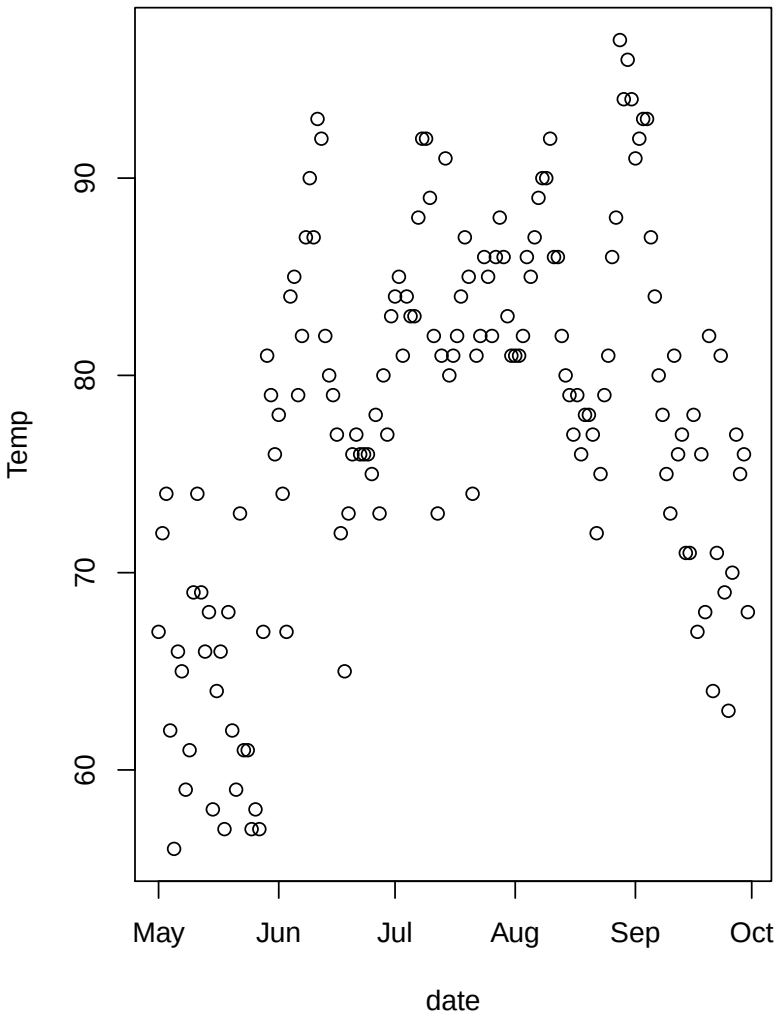
Exercise #3

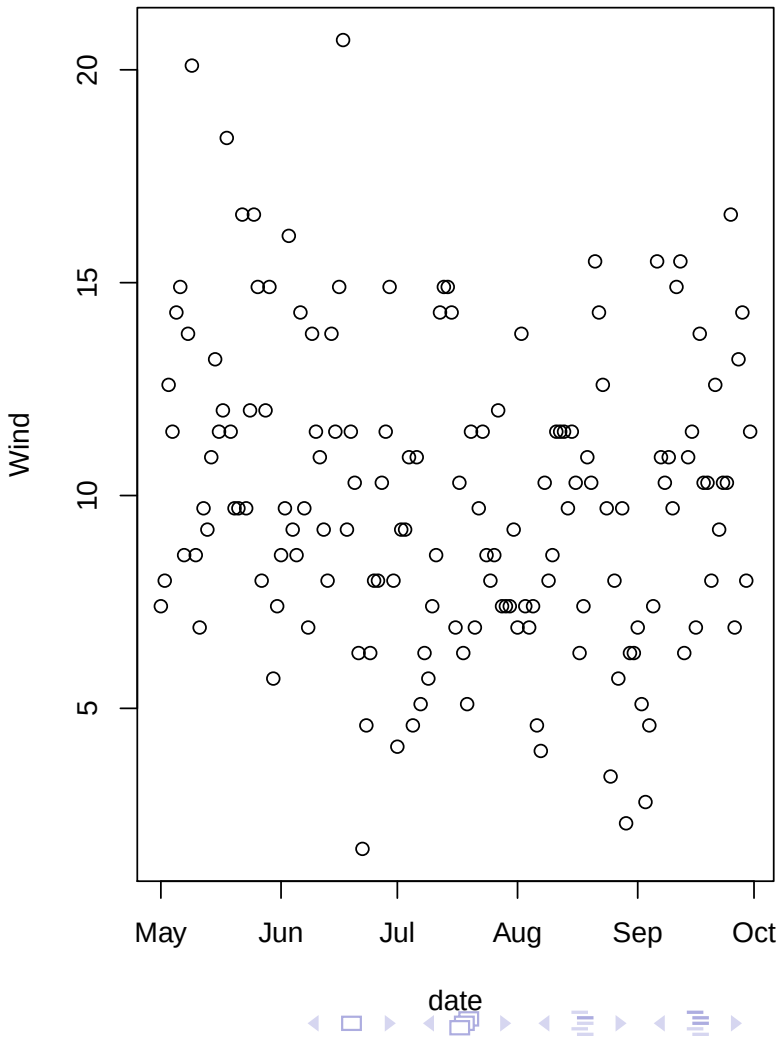
Real association?



Exercise #3

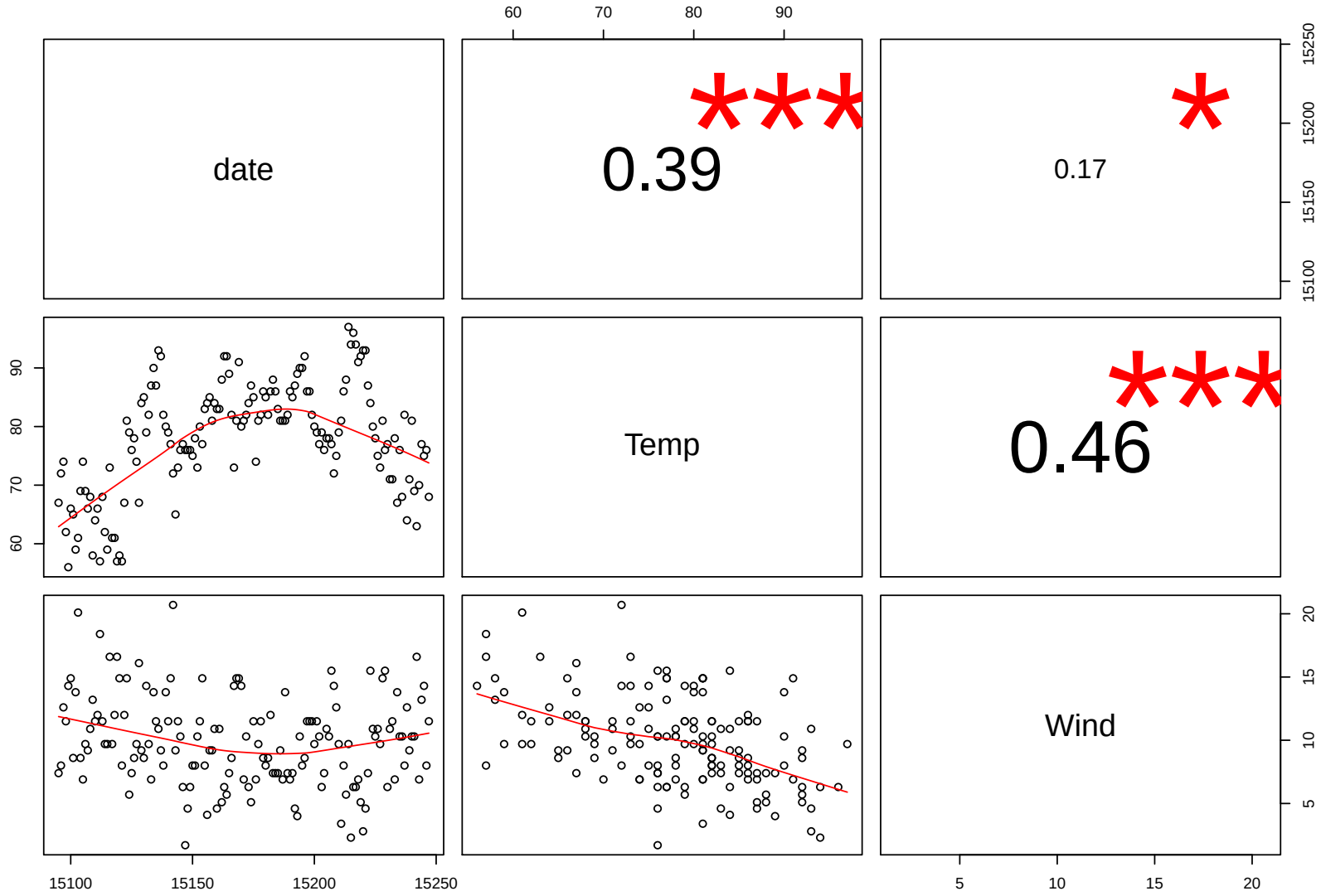
Real association?





Exercise #3

Real association?



Computation exercises

Required formulas

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad s_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

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

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{r_{XY} - r_{X\mathbf{Z}} r_{Y\mathbf{Z}}}{\sqrt{(1 - r_{X\mathbf{Z}}^2)(1 - r_{Y\mathbf{Z}}^2)}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \hat{\beta} = \frac{COV(x,y)}{VAR(x)} \quad \hat{y} = \beta x + \alpha$$

Computation exercises

Exercise 1

Compute the correlation and build linear models:

Grade	Monthly scholarship	Spent on books
3	22000 Ft	4000 Ft
4	24000 Ft	3000 Ft
5	27000 Ft	2500 Ft
3.5	24000 Ft	3500 Ft
2	23000 Ft	2000 Ft

It was a pleasure!

Gergely Daróczi
daroczi.gergely@btk.ppke.hu