Data Visualisation and Dashboarding

Week 5 – Exploratory Data Analysis

UNIVERSITY OF WESTMINSTER

“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question.”

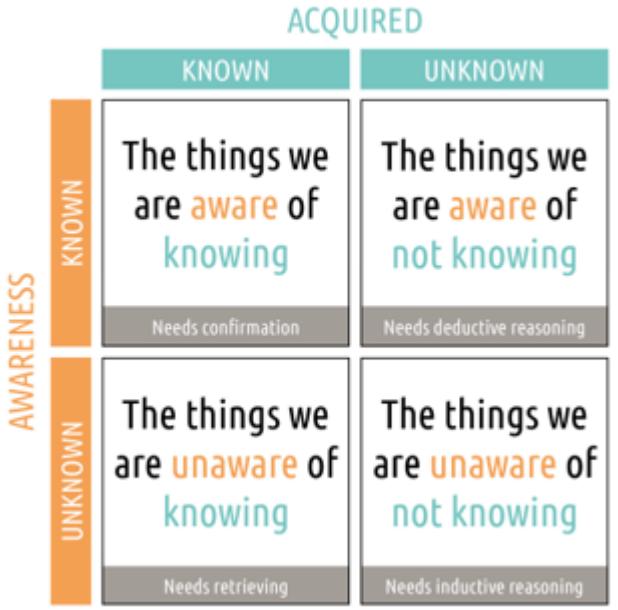
— John Tukey

What do we know?

It’s easy to learn facts we are aware they exist

It’s hard to learn facts we’re unaware of

EDA aims to discover unknown unknowns – surprising knowledge we didn’t event know to look for!



What is Exploratory Data Analysis?

Systematic exploration of data

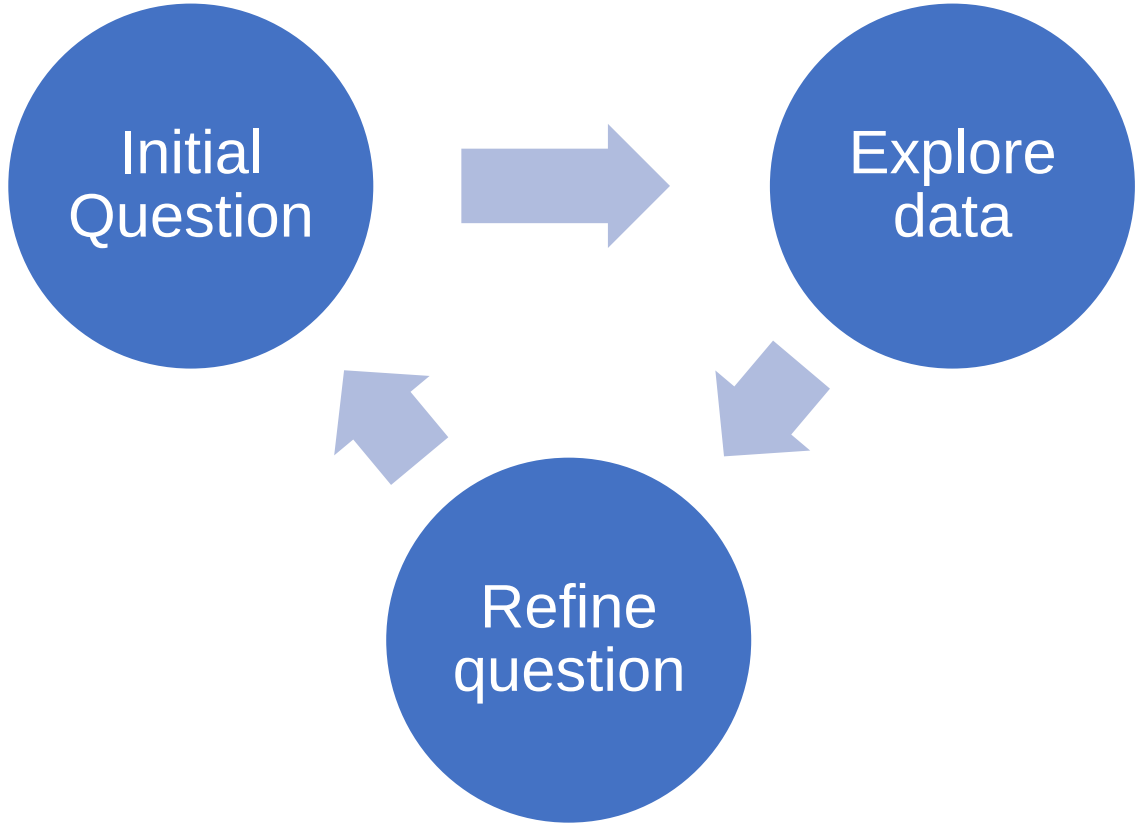
Goal: Data understanding

Not a rule based, formal process

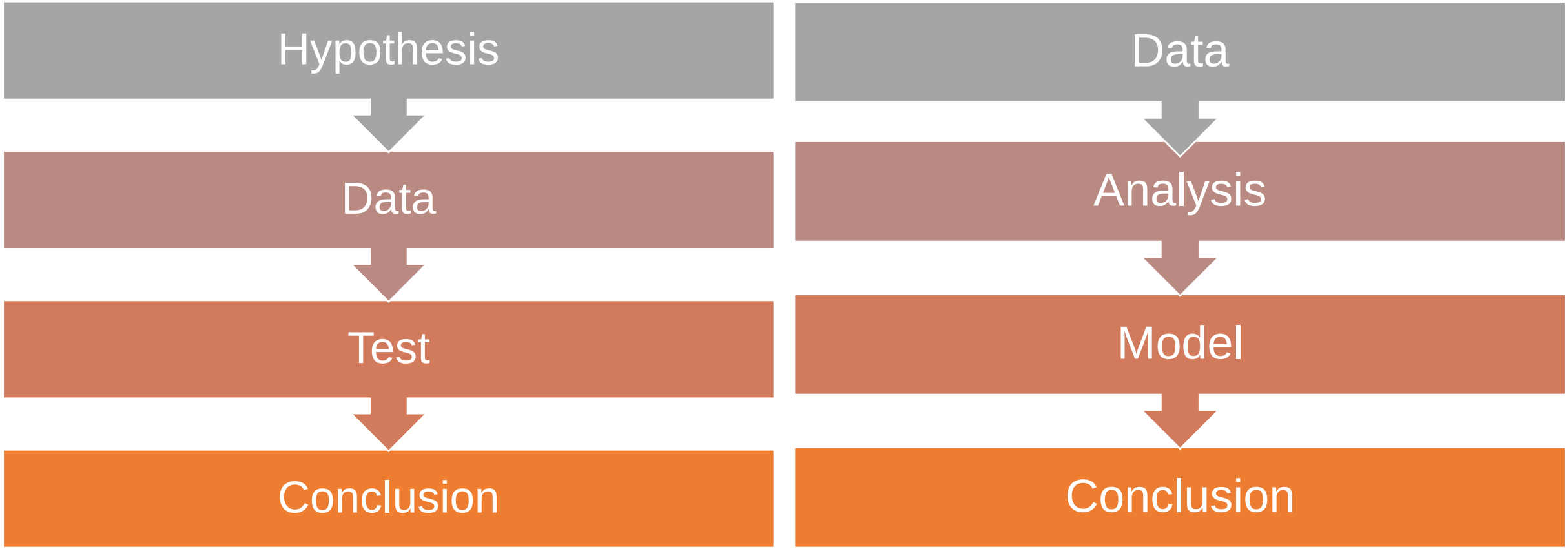
Detect data artifacts, mistakes

Check assumptions

Suggest possible statistical models



Confirmist vs Exploratory Data Analysis



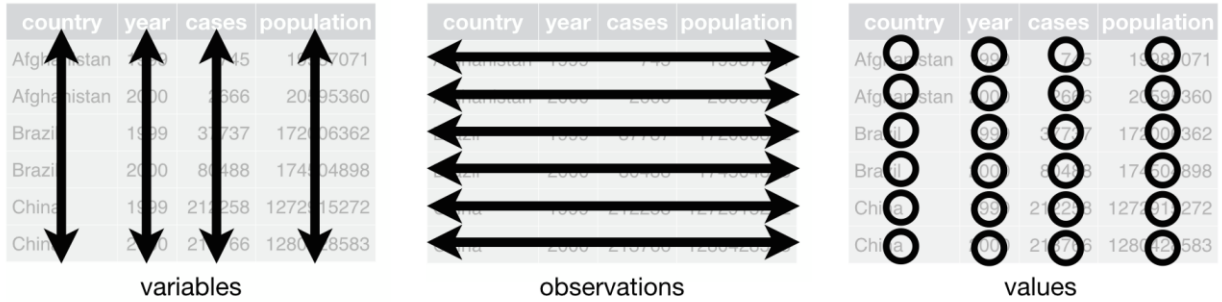
Steps of a Exploratory Data Analysis

What is tidy data?

Each variable must have its own column.

Each observation must have its own row.

Each value must have its own cell.



A variable contains all values of the same attribute (e.g. height, weight, mark).

A value is the state of the variable when you measure it.

An observation contains all values measured on the same subject (e.g. person, country, experiment) under similar condition.

Is this tidy data?

What values do we have in the tables?

What variables do we have?

	Amy	Bart	Charlie
Treatment A	16	-	3
Treatment B	2	11	1

	Treatment A	Treatment B
Amy	16	2
Bart	-	11
Charlie	3	1

Tidy data example

Variables: People, Treatment and Result

Observations are Result per each subject and treatment

Person	Treatment	Result
Amy	A	16
Bart	A	-
Charlie	A	3
Amy	B	2
Bart	B	11
Charlie	B	1

Independent and dependent variables

Independent variables

- Explanatory variables
- Predictor variables
- Right-hand-side variables
- Often (not always) categorical
- Answers: Who? What? When? Where?

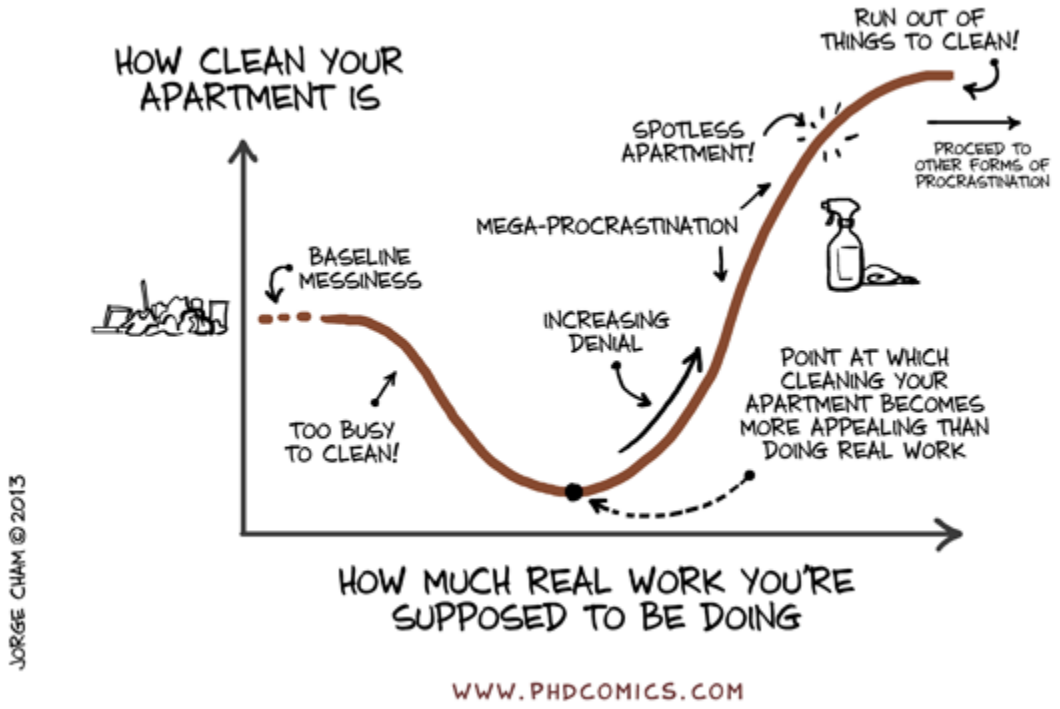
Dependent variables

- Response variables
- Outcomes variables
- Left-hand-side variables
- Often (not always) continuous
- Answers: How much? How many? How long?

$y = ax + b$ (y is the dependent variable)

Independent		Dependent
Person	**Treatment**	**Result**
Amy	A	16
Bart	A	-
Charlie	A	3
Amy	B	2
Bart	B	11
Charlie	B	1

Which one is the dependent variable?



Which variables are independent/dependent?

Independent

Dependent

Period and Financial year	Reporting Period	Days in period	Period beginning	Period ending	Bus journeys (m)	Underground journeys (m)	DLR Journeys (m)	Tram Journeys (m)	Overground Journeys (m)	Emirates Airline Journeys (m)	TfL Rail Journeys (m)
01_10/11	1	31	01-Apr-10	01-May-10	189.1	90.5	6.3	2.3			
02_10/11	2	28	02-May-10	29-May-10	181.6	84.5	5.8	2.2			
03_10/11	3	28	30-May-10	26-Jun-10	175.9	84.3	5.8	2.1			
04_10/11	4	28	27-Jun-10	24-Jul-10	183.4	86.5	6.1	2.1			
05_10/11	5	28	25-Jul-10	21-Aug-10	160.4	82.9	5.8	2.0			
06_10/11	6	28	22-Aug-10	18-Sep-10	175.8	80.9	5.5	2.0			
07_10/11	7	28	19-Sep-10	16-Oct-10	189.8	88.7	6.3	2.3			
08_10/11	8	28	17-Oct-10	13-Nov-10	179.9	90.3	6.7	2.2	5.6		
09_10/11	9	28	14-Nov-10	11-Dec-10	178.8	90.6	6.4	2.3	5.4		
10_10/11	10	28	12-Dec-10	08-Jan-11	140.1	72.5	4.8	1.8	3.5		
11_10/11	11	28	09-Jan-11	05-Feb-11	183.0	84.4	6.3	2.1	5.2		
12_10/11	12	28	06-Feb-11	05-Mar-11	177.2	87.8	6.5	2.2	5.2		
13_10/11	13	26	06-Mar-11	31-Mar-11	173.9	83.5	6.0	2.1	5.8		
01_11/12	1	30	01-Apr-11	30-Apr-11	183.8	91.2	6.4	2.1	6.3		
02_11/12	2	28	01-May-11	28-May-11	186.1	87.8	6.3	2.2	6.4		
03_11/12	3	28	29-May-11	25-Jun-11	181.7	88.9	6.1	2.2	6.9		
04_11/12	4	28	26-Jun-11	23-Jul-11	186.7	92.5	6.4	2.3	7.6		
05_11/12	5	28	24-Jul-11	20-Aug-11	161.1	85.5	6.2	1.9	7.7		
06_11/12	6	28	21-Aug-11	17-Sep-11	173.9	85.3	6.4	2.1	7.7		
07_11/12	7	28	18-Sep-11	15-Oct-11	193.4	93.1	7.5	2.4	8.7		
08_11/12	8	28	16-Oct-11	12-Nov-11	185.2	95.8	7.3	2.3	8.8		
09_11/12	9	28	13-Nov-11	10-Dec-11	189.4	97.1	7.1	2.4	9.0		
10_11/12	10	28	11-Dec-11	07-Jan-12	151.2	79.3	5.2	1.9	6.9		
11_11/12	11	28	08-Jan-12	04-Feb-12	181.4	89.8	6.9	2.2	8.7		
12_11/12	12	28	05-Feb-12	03-Mar-12	179.5	91.5	7.0	2.0	8.8		
13_11/12	13	28	04-Mar-12	31-Mar-12	191.2	92.7	7.3	2.3	9.1		

Qualitative and Quantitative Scales

Qualitative		Quantitative	
Nominal	**Ordinal**	**Interval**	**Ratio**
Labelled data	**Ordered labels**	**Quantitative data on relative scale**	**Quantitative data on absolute scale**
Alice, Bob, Chris,…	Gold, Silver, Bronze	Constant interval	Constant interval
Bexley, Bromley, Camden, Croydon,…	Excellent, Good, Poor	No true zero	True zero
	January, February, March, …	Temperature in Celsius	Temperature in Kelvin
		Profit	Revenue

Variation

Variation

Tendency of a variable to change

Only looking at one variable at a time

Frequency distribution: how many time does a value (or a range of values) appear?

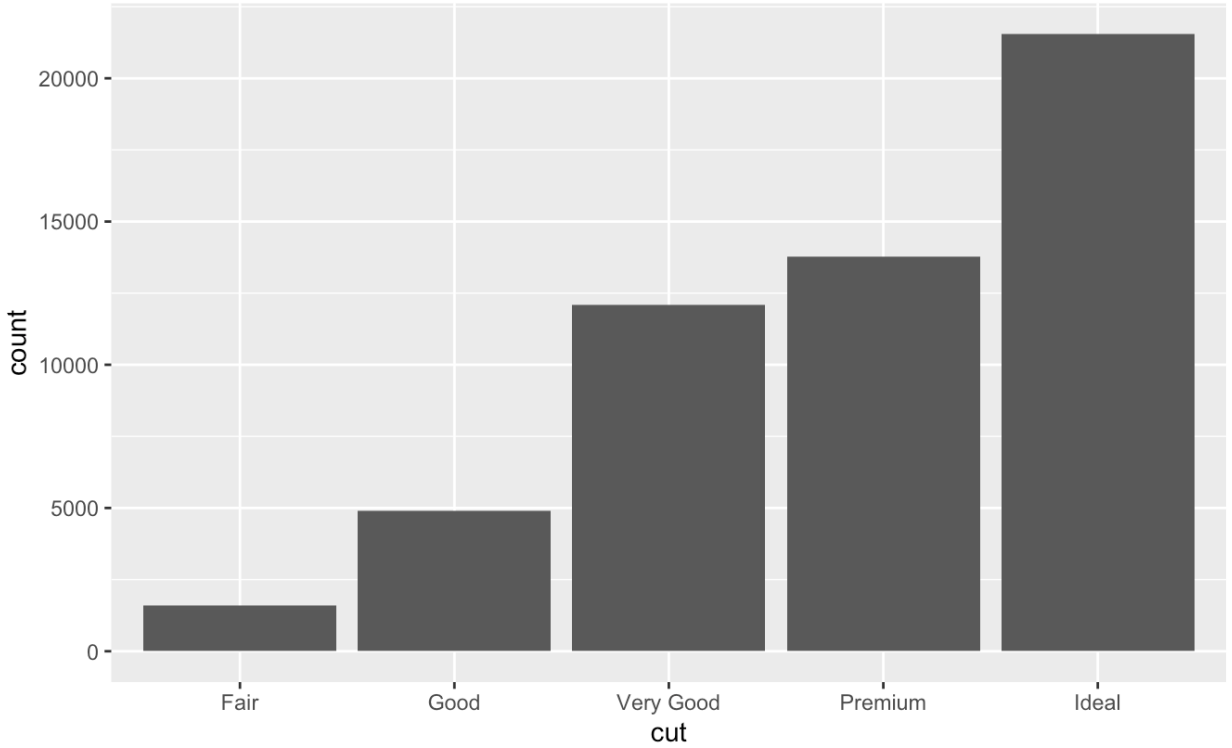
UNIVERSITY OF WESTMINSTER

Analysing the diamonds dataset

50000 round cut diamonds

Price, carat, dimensions, cut quality, colour quality etc.

Use bar chart to show distribution of categorical variable (cut)

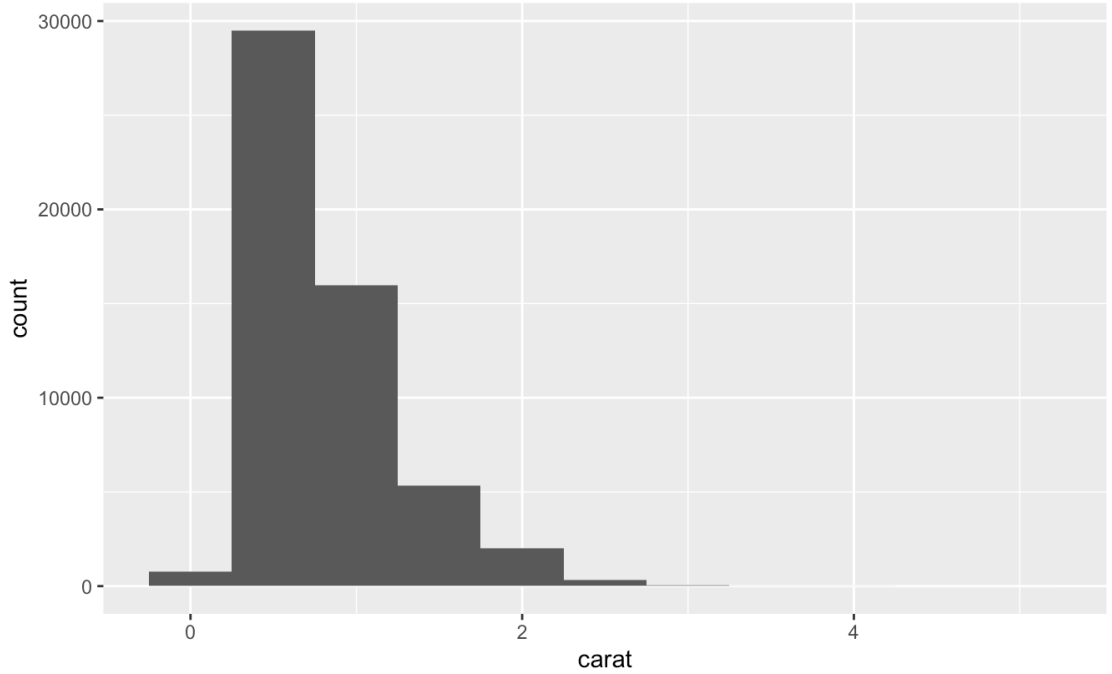


Histogram

Shows distribution of continuous variables

Similar values are grouped into “bins”

Number of bins / width of bins is important!



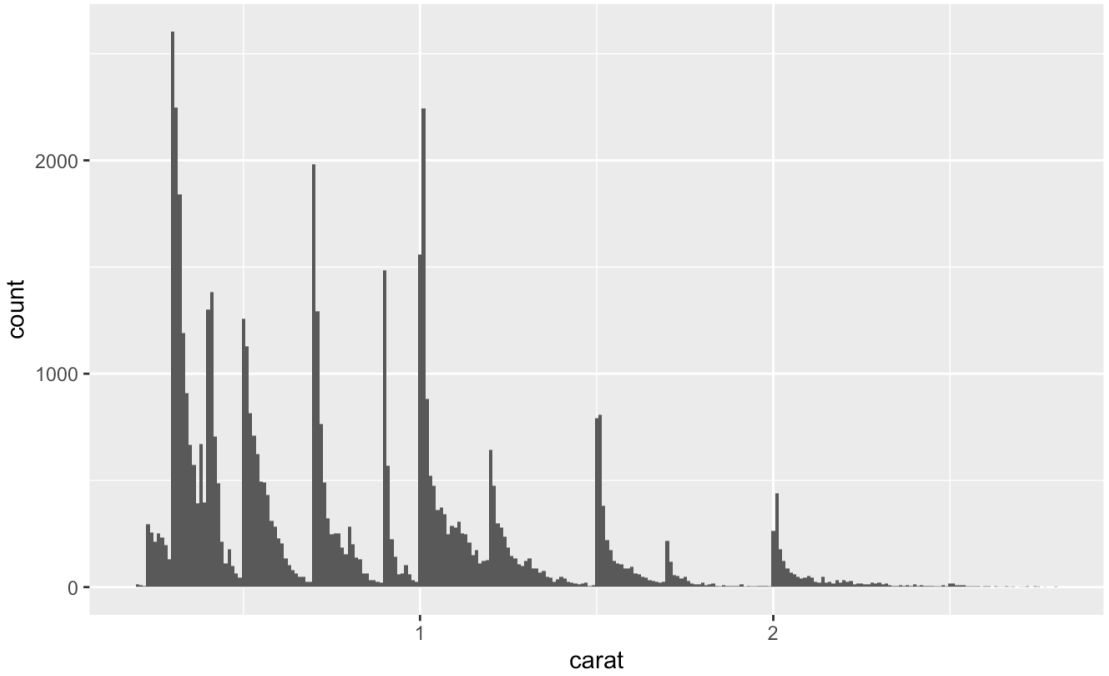
Histogram

Using smaller bins can reveal interesting details about a data set

There are more diamonds at the at whole carats and common fraction of carats

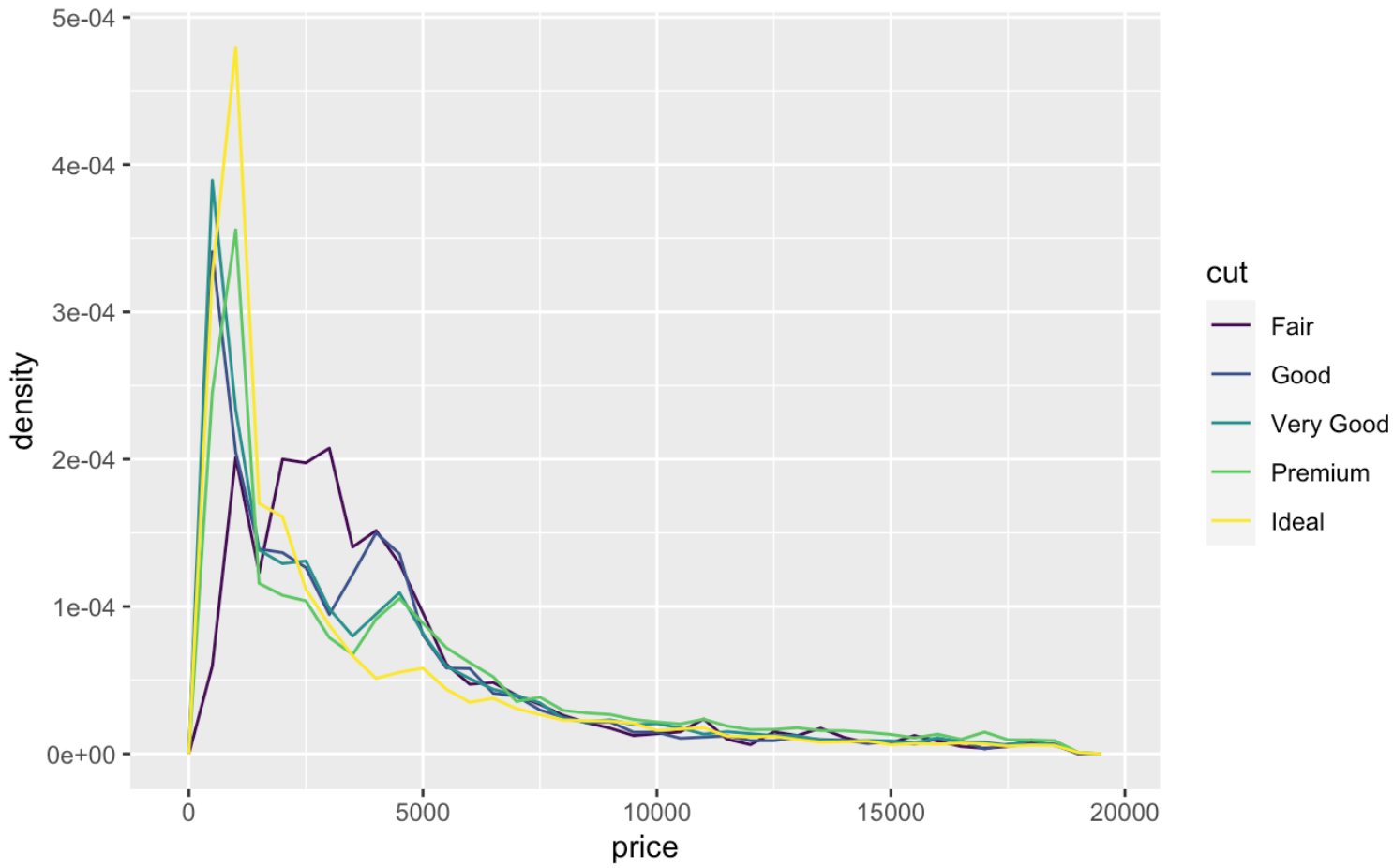
More diamonds slightly to the right of each peak

No diamonds bugger than 3 carat



Frequency Polygon

Useful to compare multiple categories

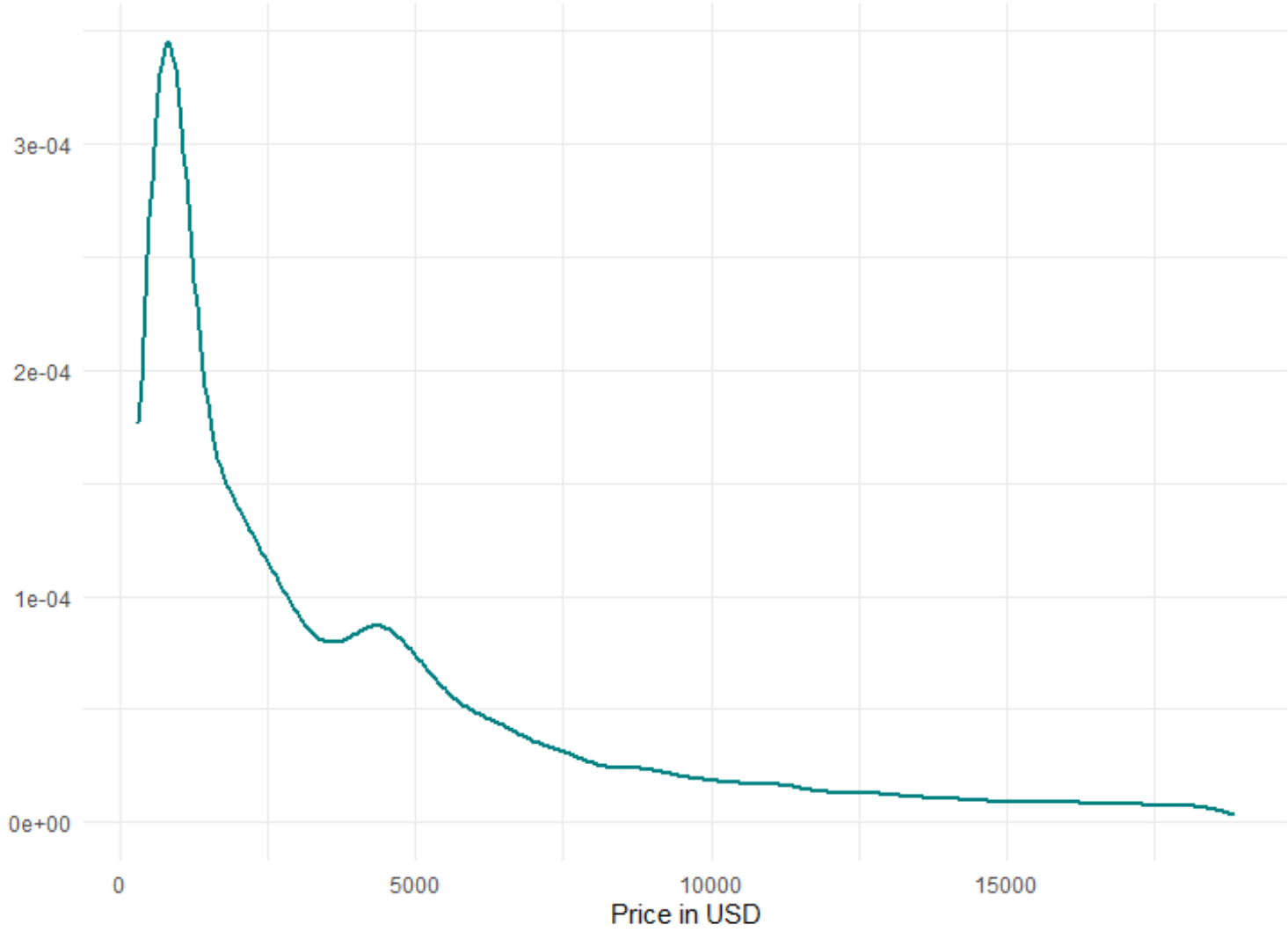


Density plot

Draws density estimate

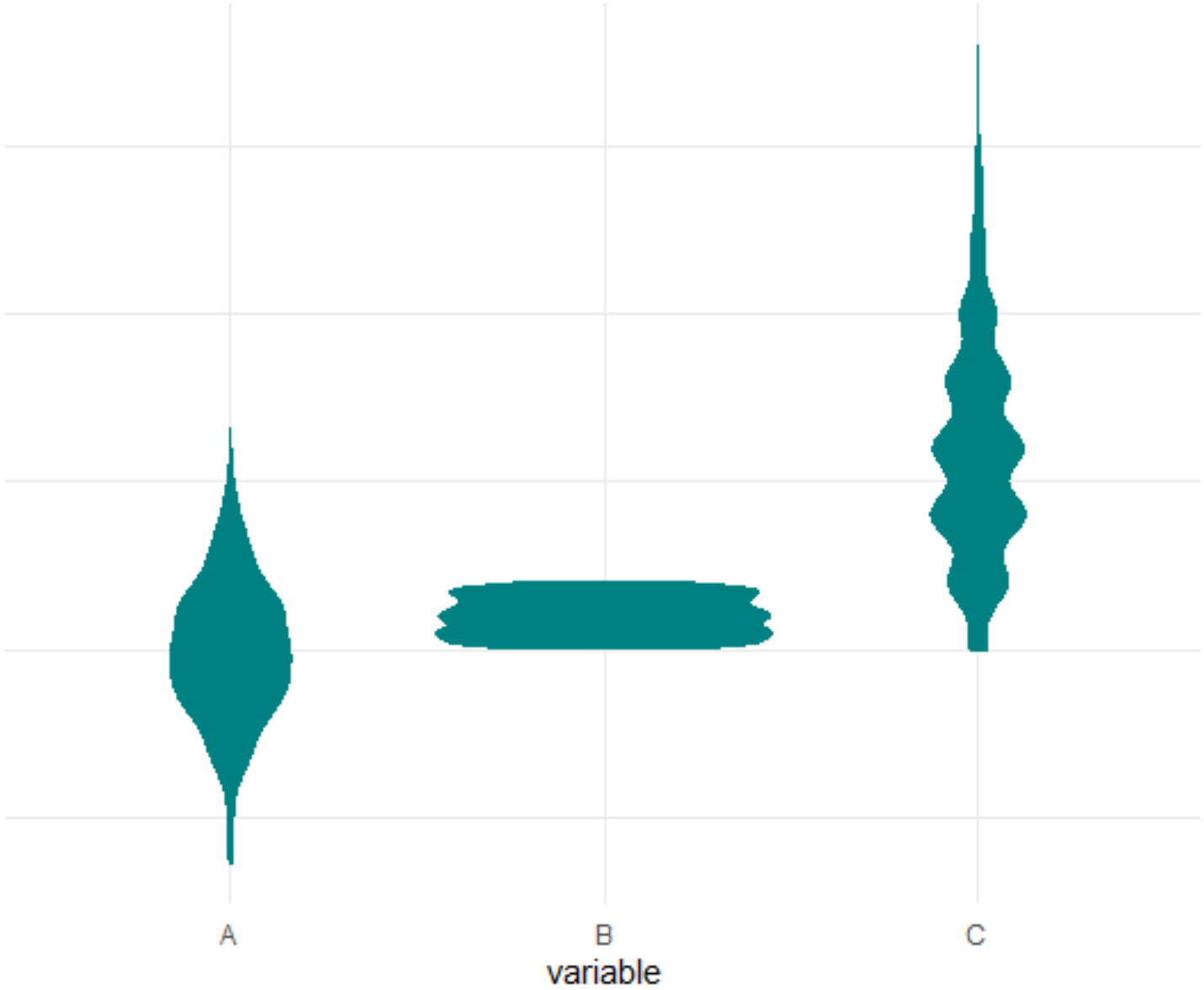
Think of as smoothed histogram

Shows relative frequency, **not** absolute frequency



Violin plot

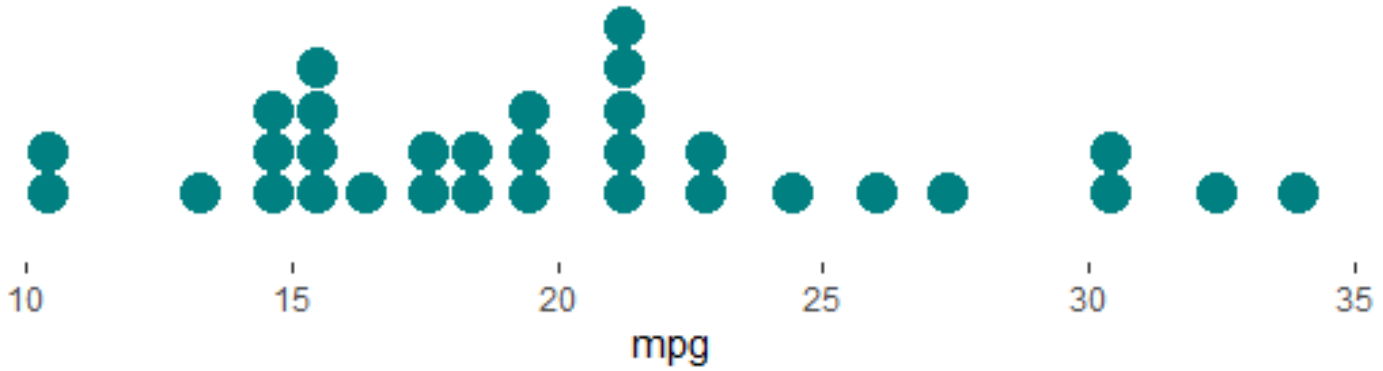
Shows mirrored density for one or more categories



Dot plot

Defined by Wilkinson, 1999.

Similar to Histogram, but easier to read exact values



Box plot

Invented by John Tukey

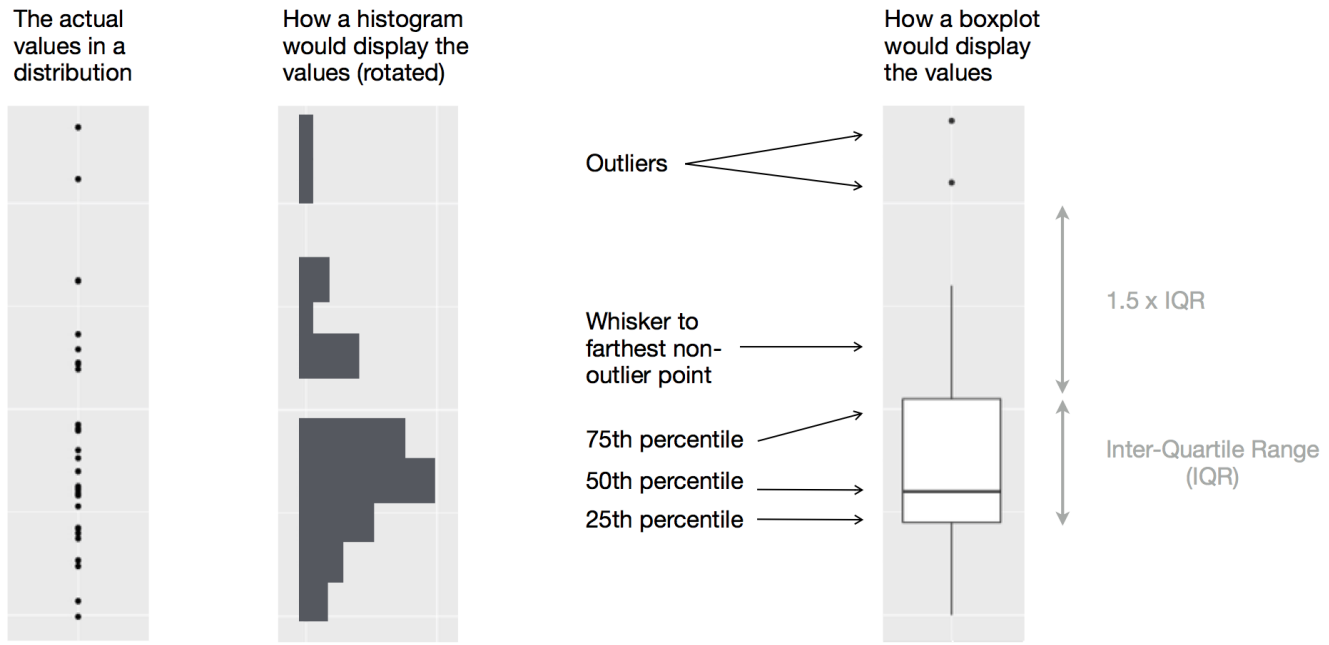
Minimum/maximum (excluding outliers)

1st quartile (Q1), median and 3rd quartile (Q3)

Interquartile range $IQR = Q3 - Q1$

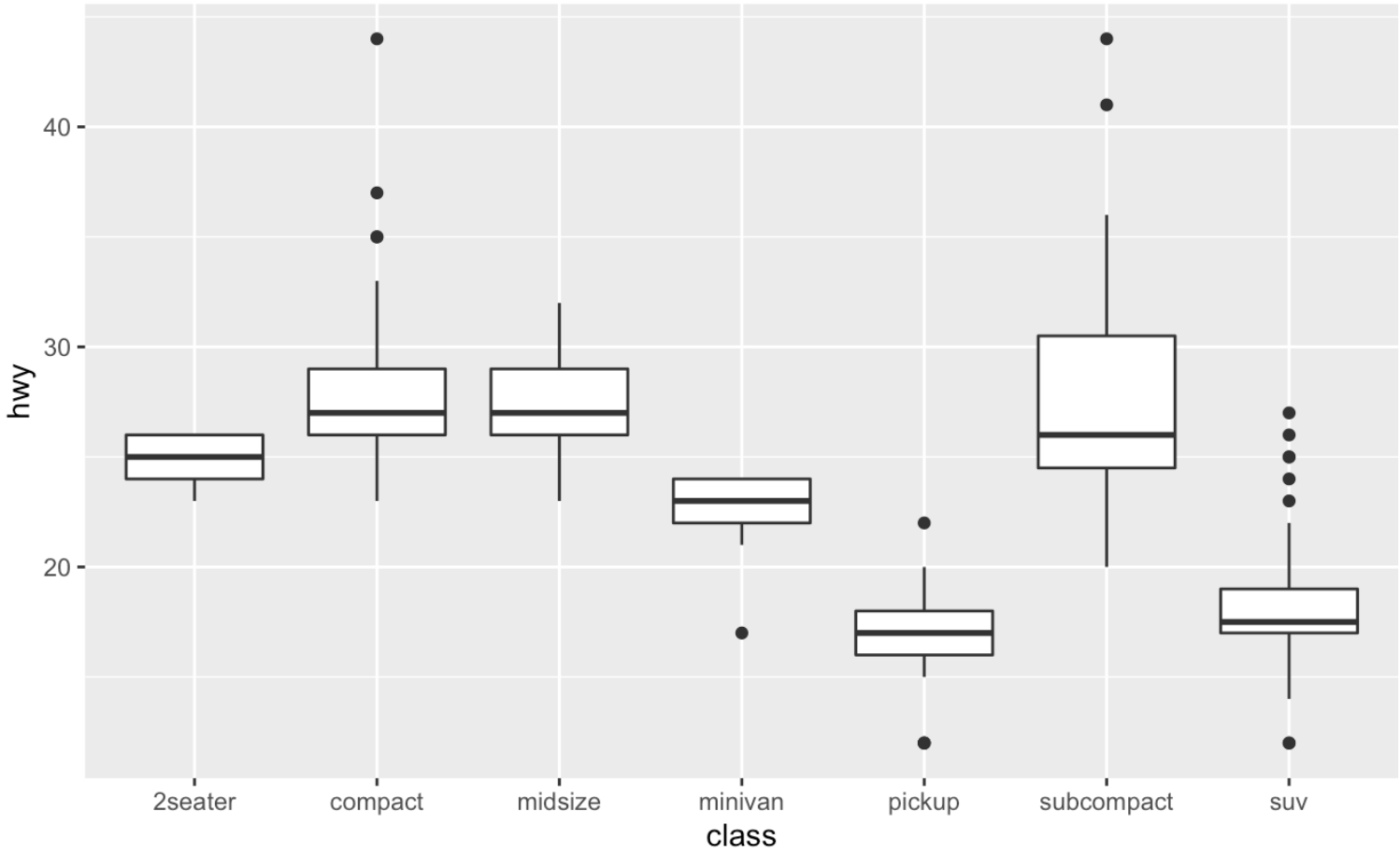
Outliers $0 < Q1 - 1.5 \cdot IQR$ and $0 > Q3 + 1.5 \cdot IQR$

Does **NOT** show average/mean!



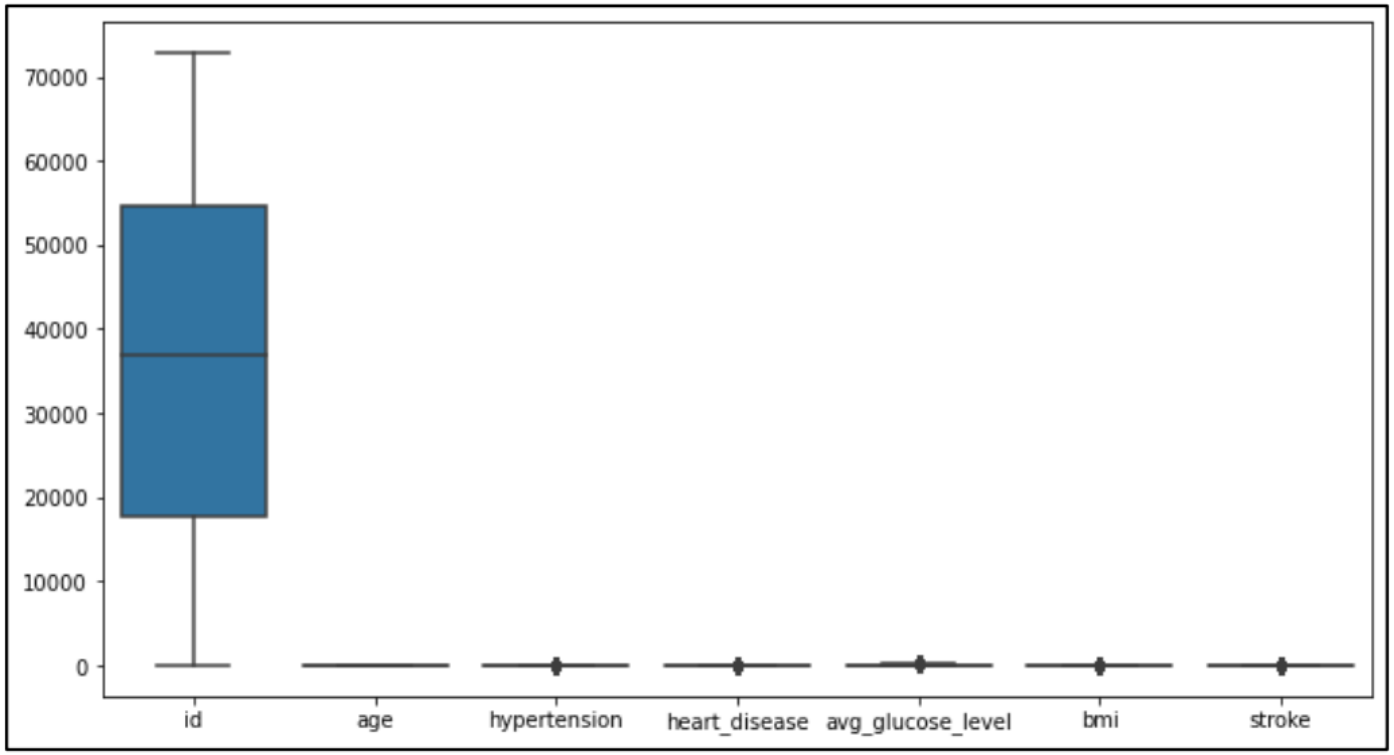
Box plot

Box plot works well to plot multiple categories of the same variable

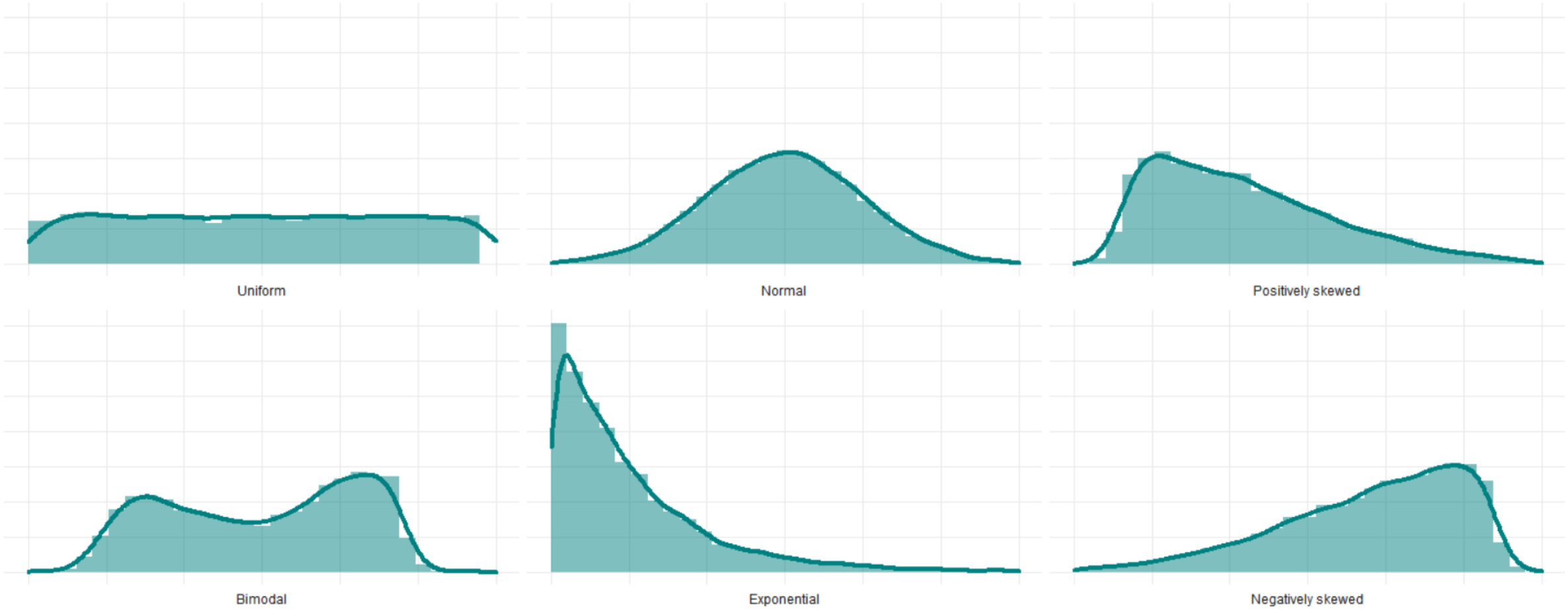


Box plot

Box plot often does **NOT** work well to plot multiple variables!



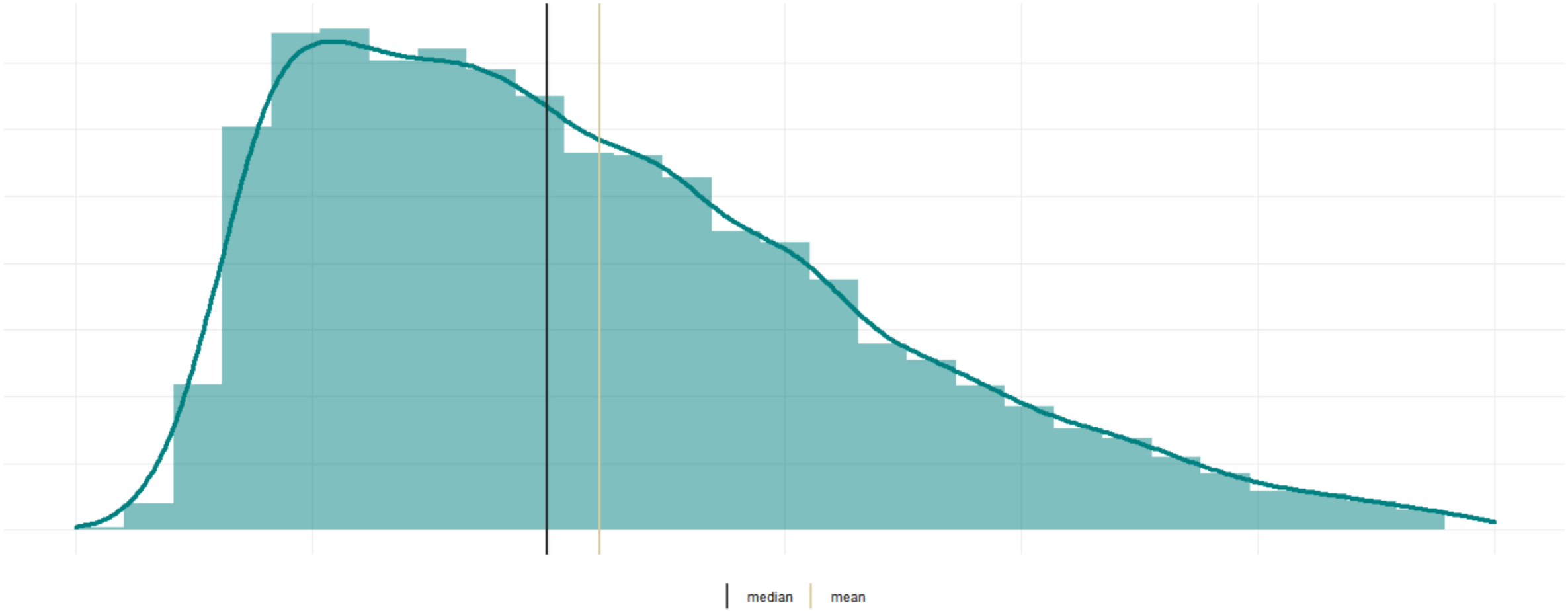
A few distributions



Implications

Uniform	Normal	Positively skewed	Negatively skewed	Bimodal	Exponential
• Range could be important	• Average and median are same • Can be described by average and standard deviation	• Mean is bigger than median	• Mean is smaller then median	• Average/median may not be very useful • Data might be better off grouped	• Could be displayed on logarithmic scale • Outliers may be important

Average and median



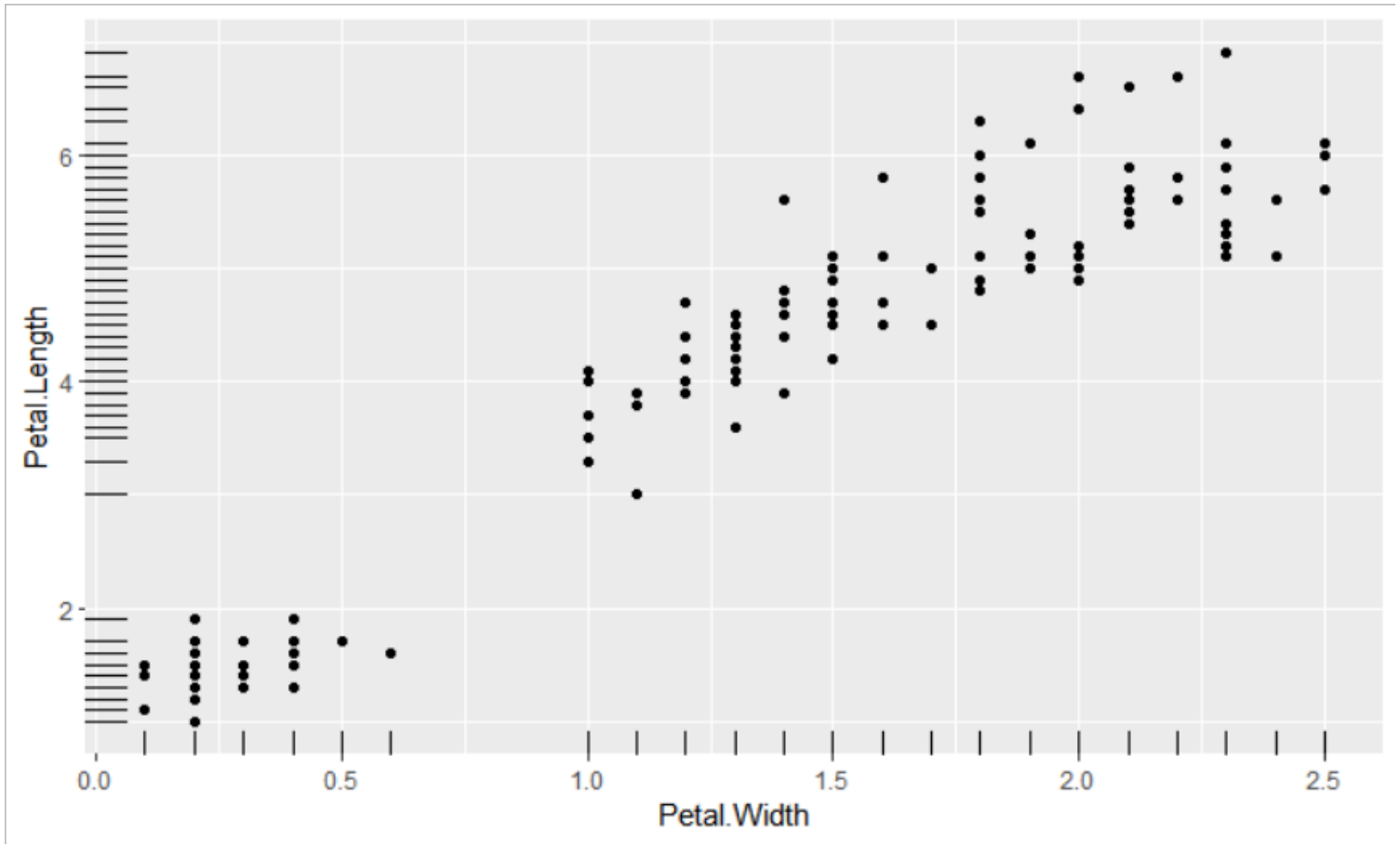
UNIVERSITY OF WESTMINSTER

Covariation

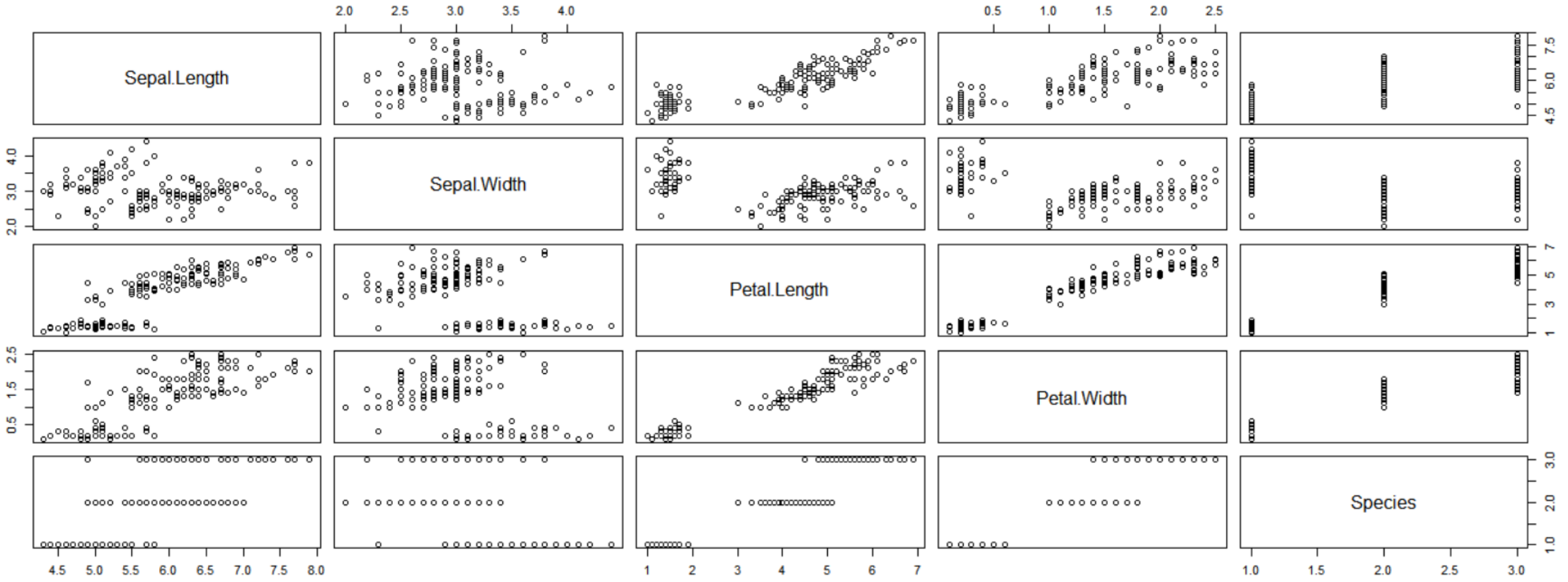
UNIVERSITY OF WESTMINSTER

Bivariate/multivariate exploration

What relationship exist between variables?

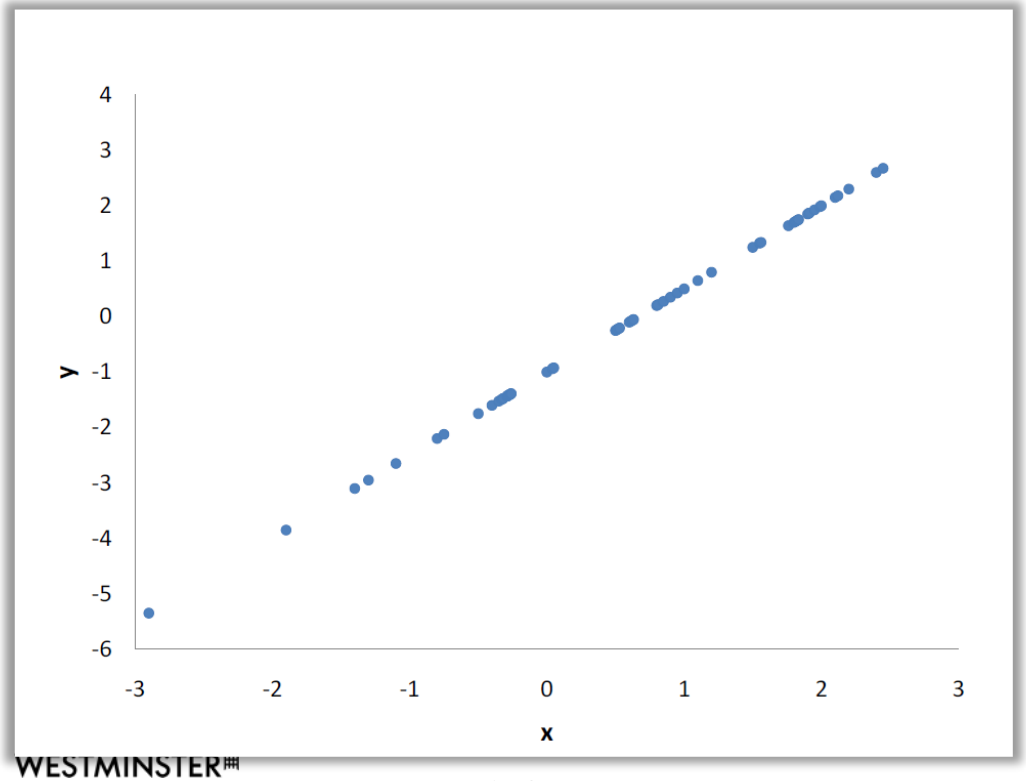


Scatter plots

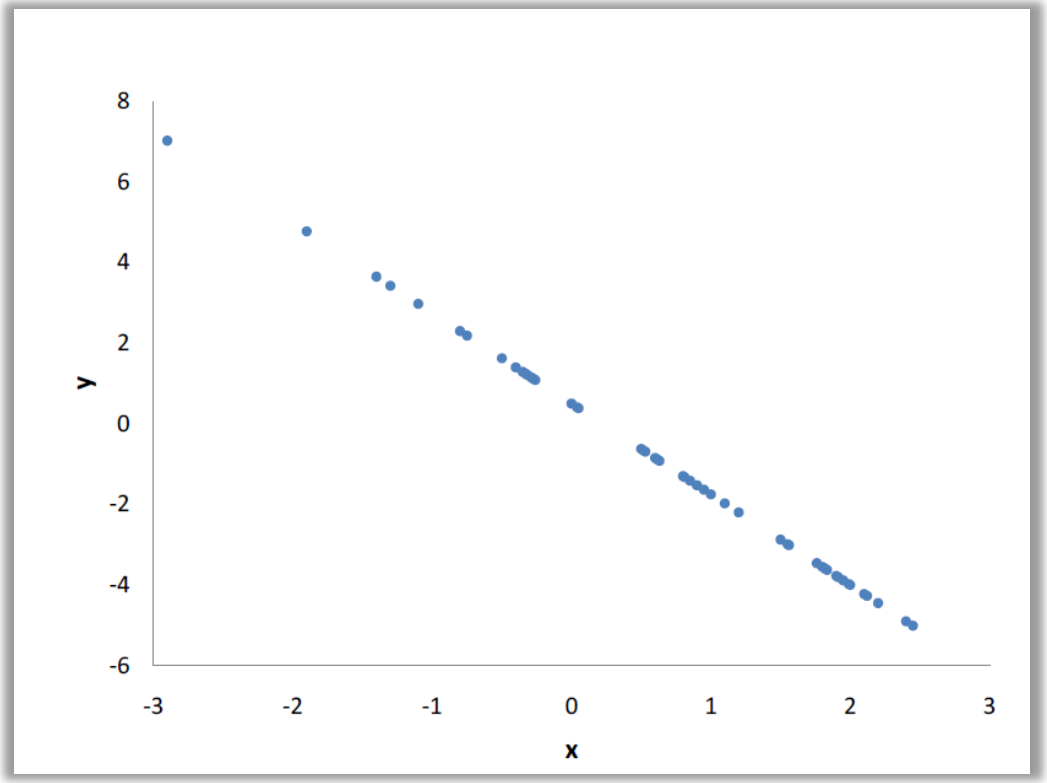


Correlation

Pearson's Correlation coefficient (r or ρ) measures the strength and direction of the linear relationship between two quantitative variables



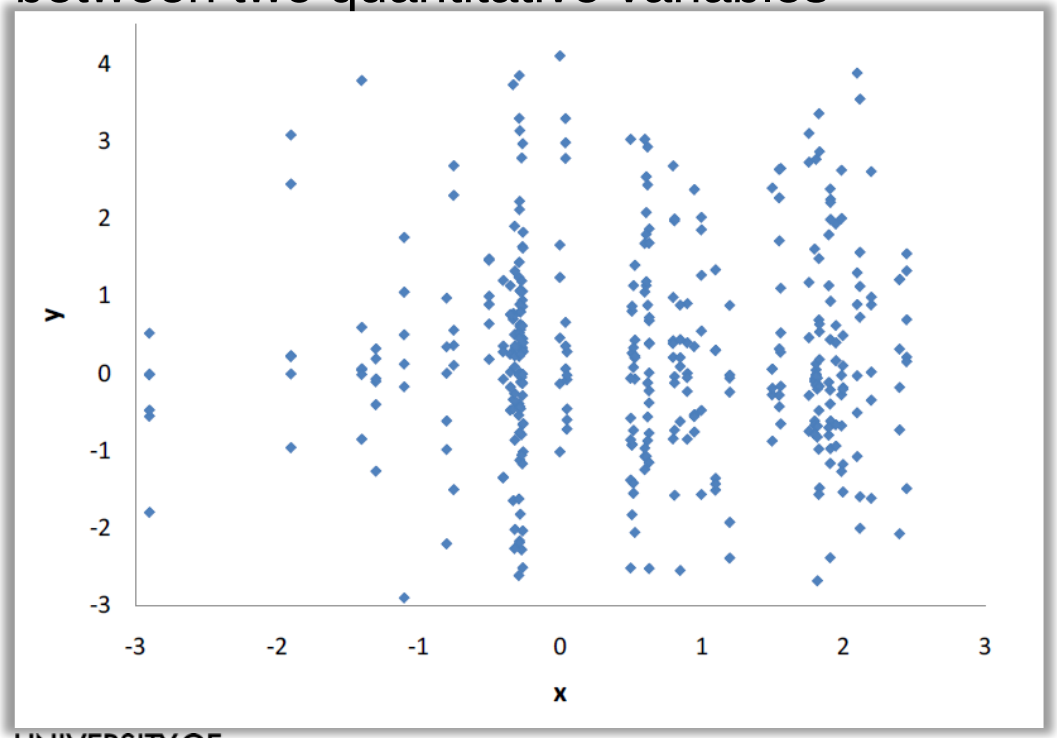
r = 1.0



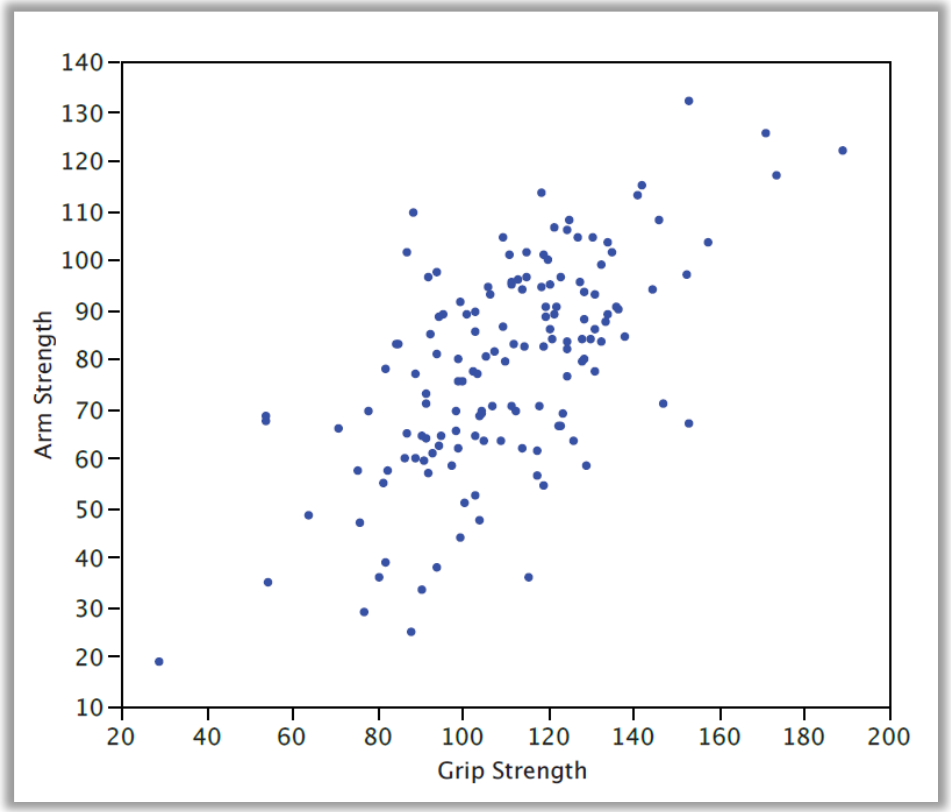
r = −1.0

Correlation

Pearson's Correlation coefficient (r or ρ) measures the strength and direction of the linear relationship between two quantitative variables

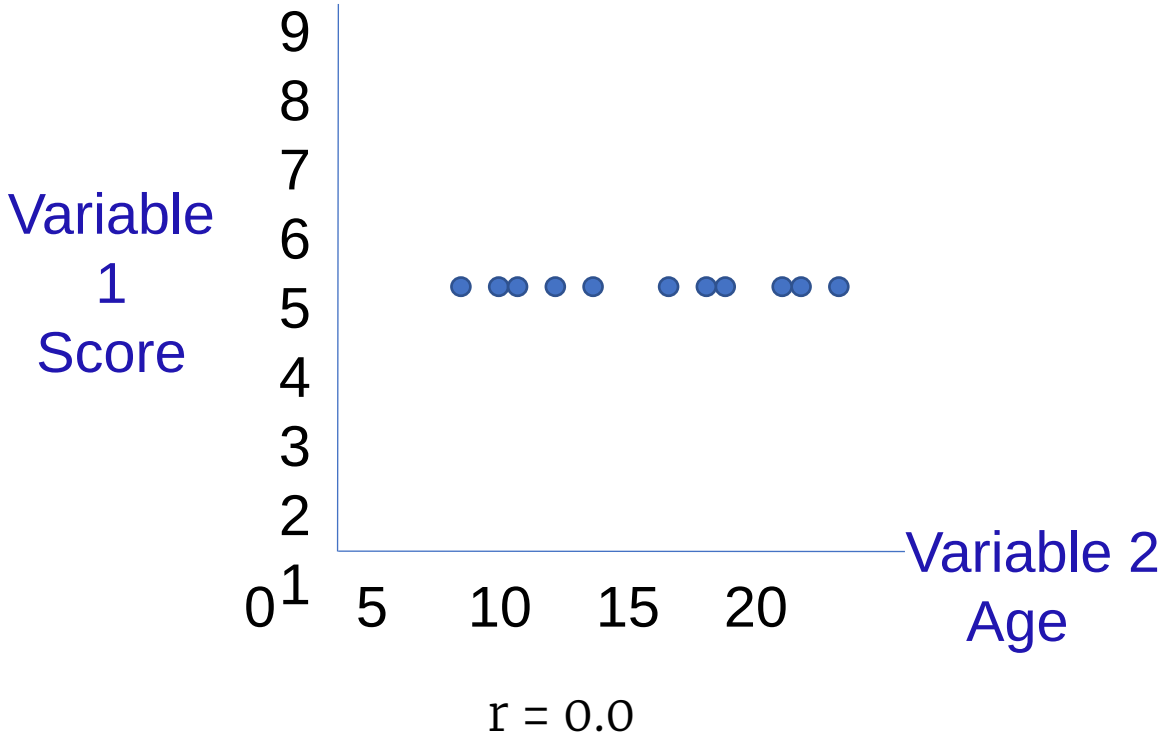


r = 0.0

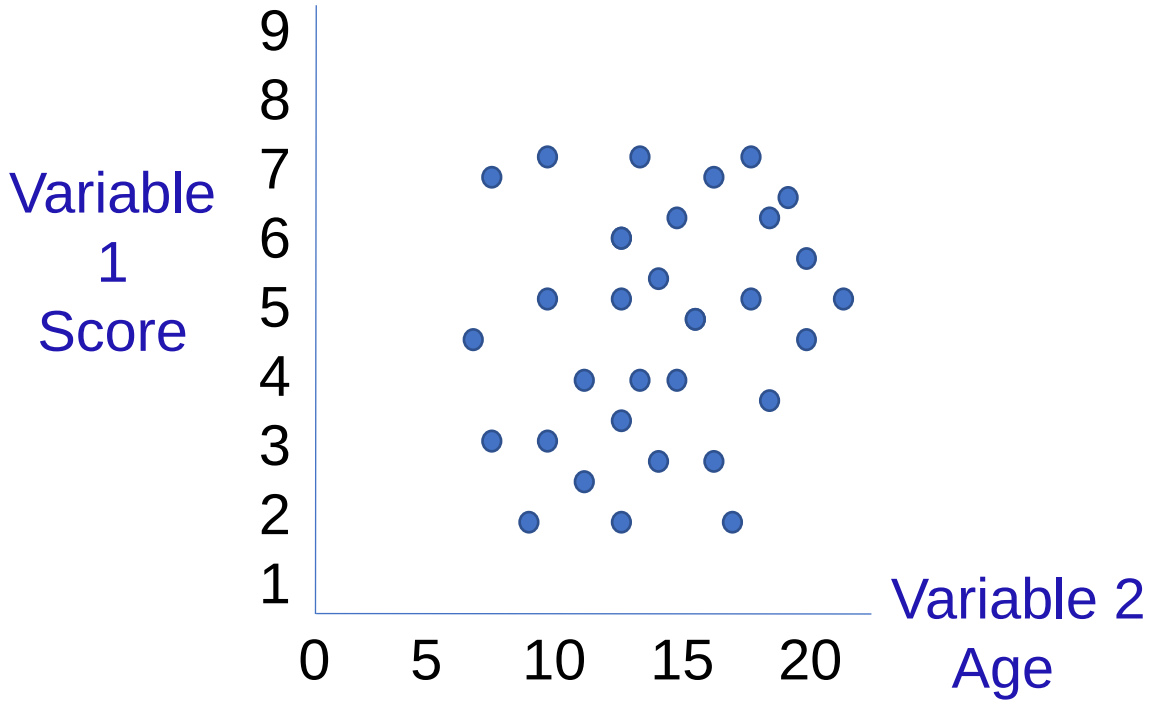


r = 0.63

Correlation



r = 0.0



r = 0.0

Strength of a Correlation

1.00 – 0.90 : very strong correlation

0.89 – 0.70 : strong correlation

0.69 – 0.40 : modest correlation

0.39 – 0.20 : weak correlation

0.19 – 0.00 : very weak correlation or no correlation

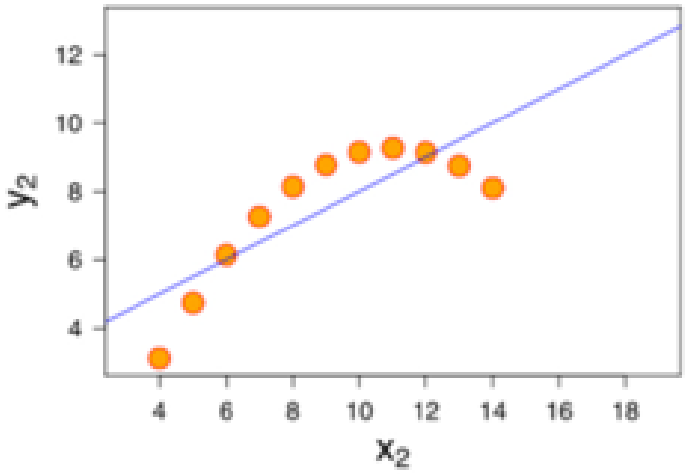
+ or – will denote direction

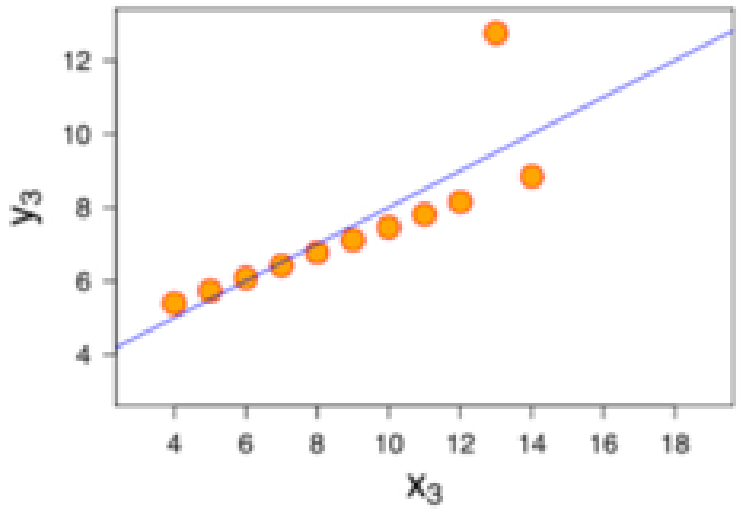
Correlation conditions

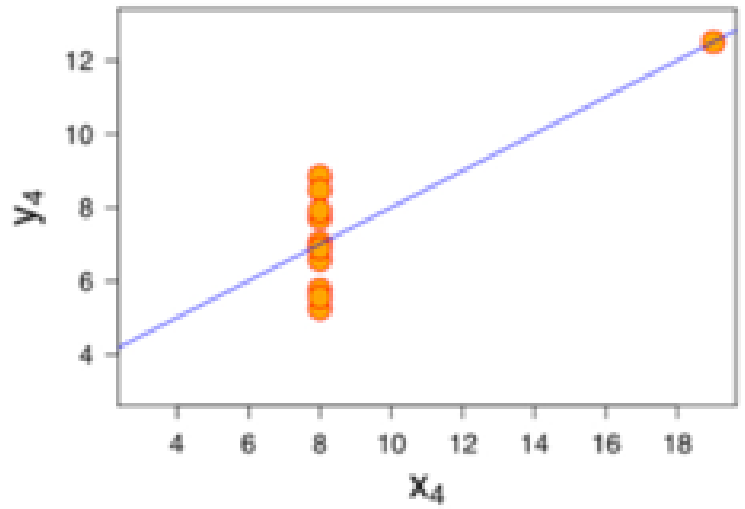
Quantitative variables

Must be linear (always look at the scatter plot before running correlation)

No outliers

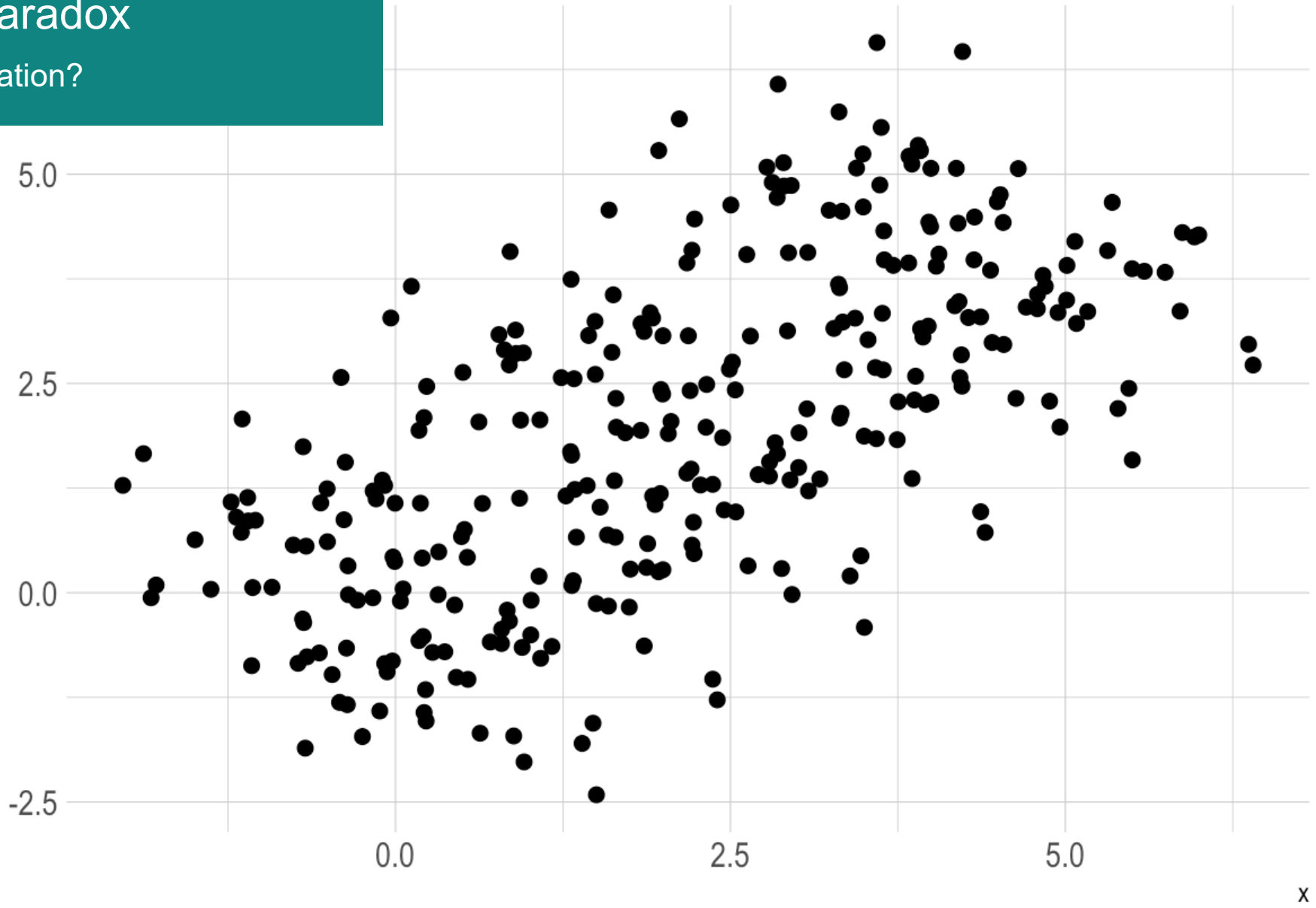






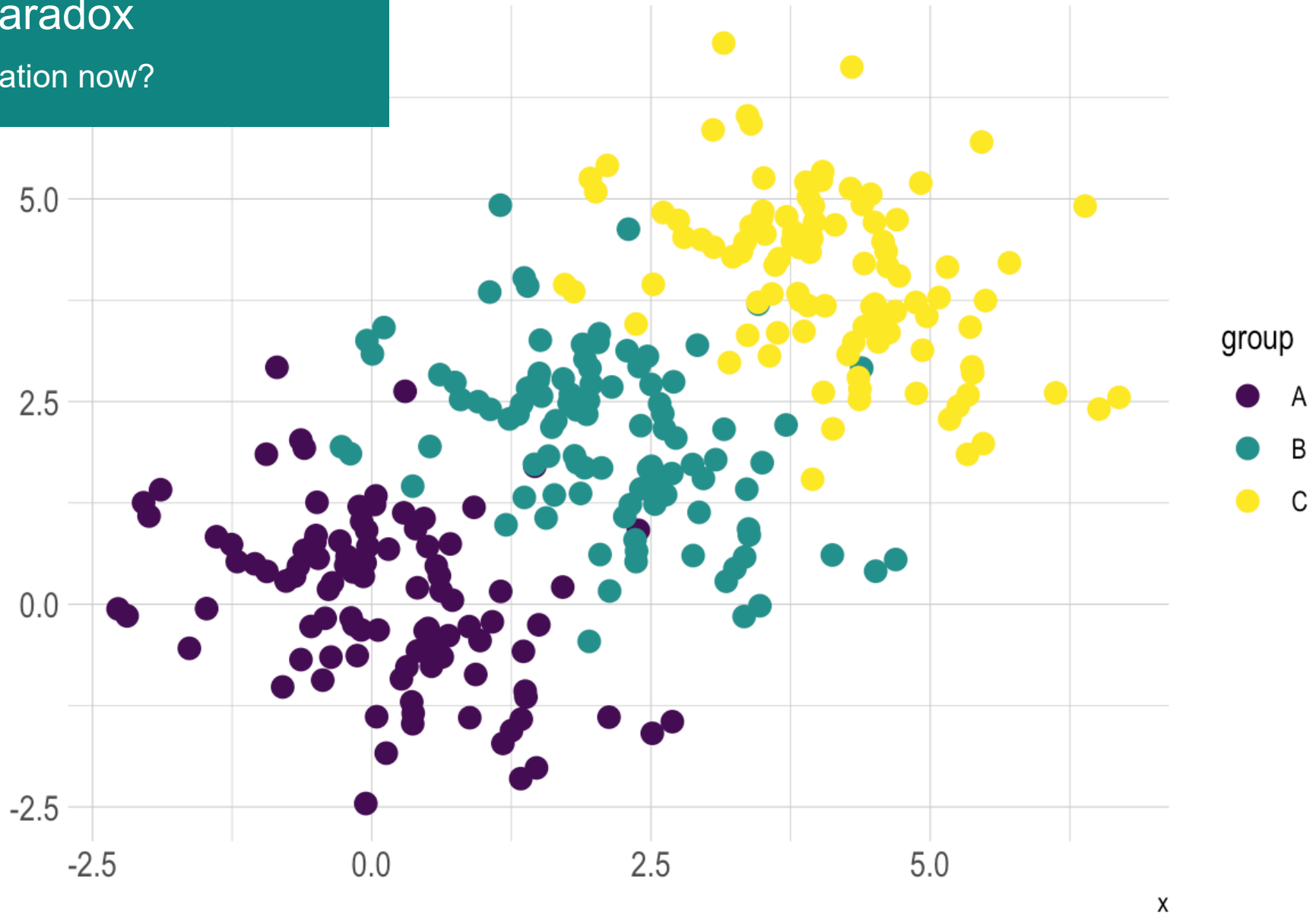
Simpson’s paradox

What’s the correlation?

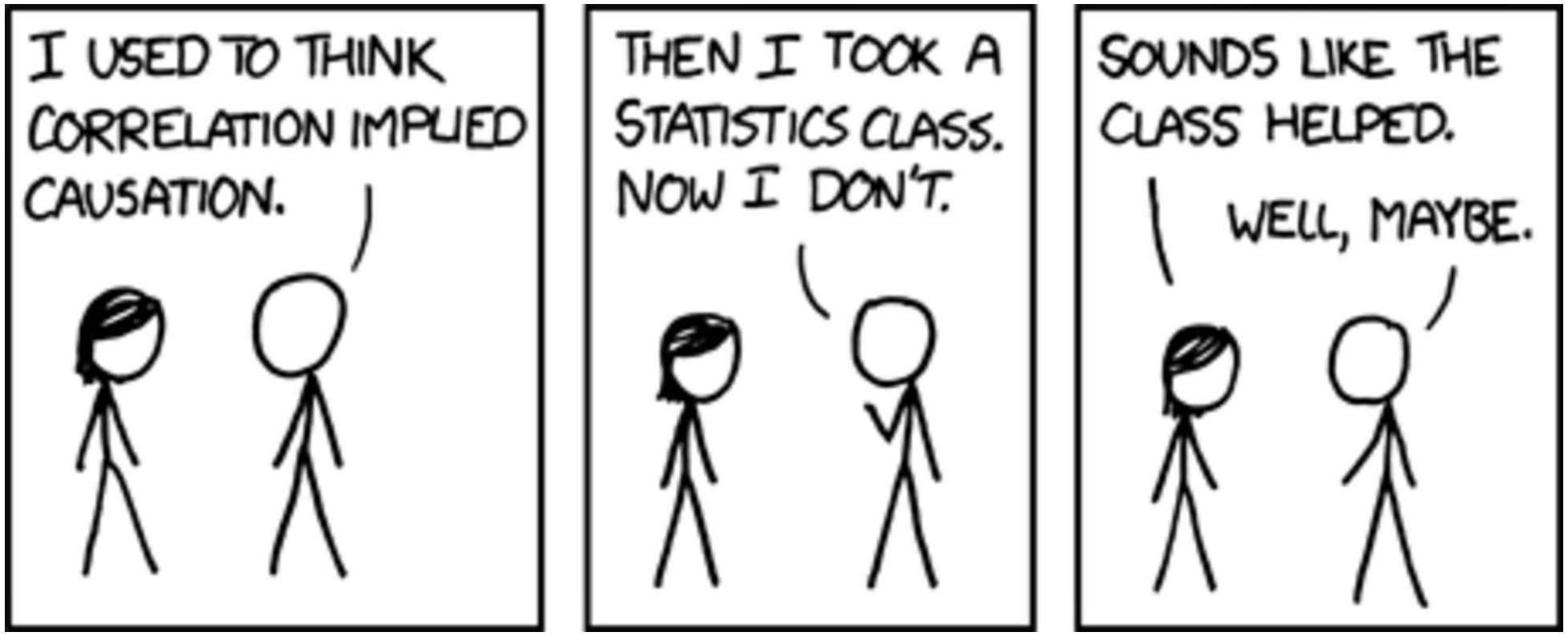


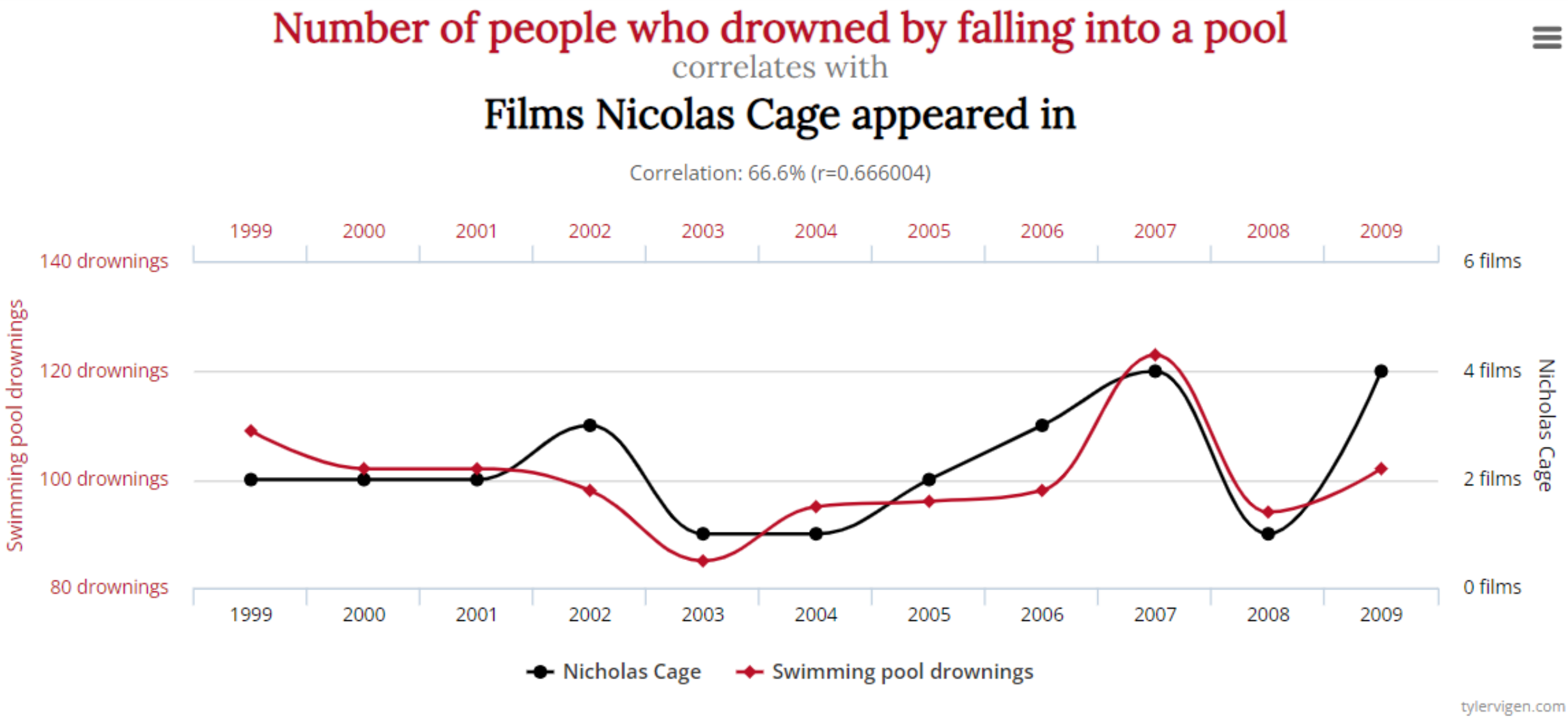
Simpson’s paradox

What’s the correlation now?

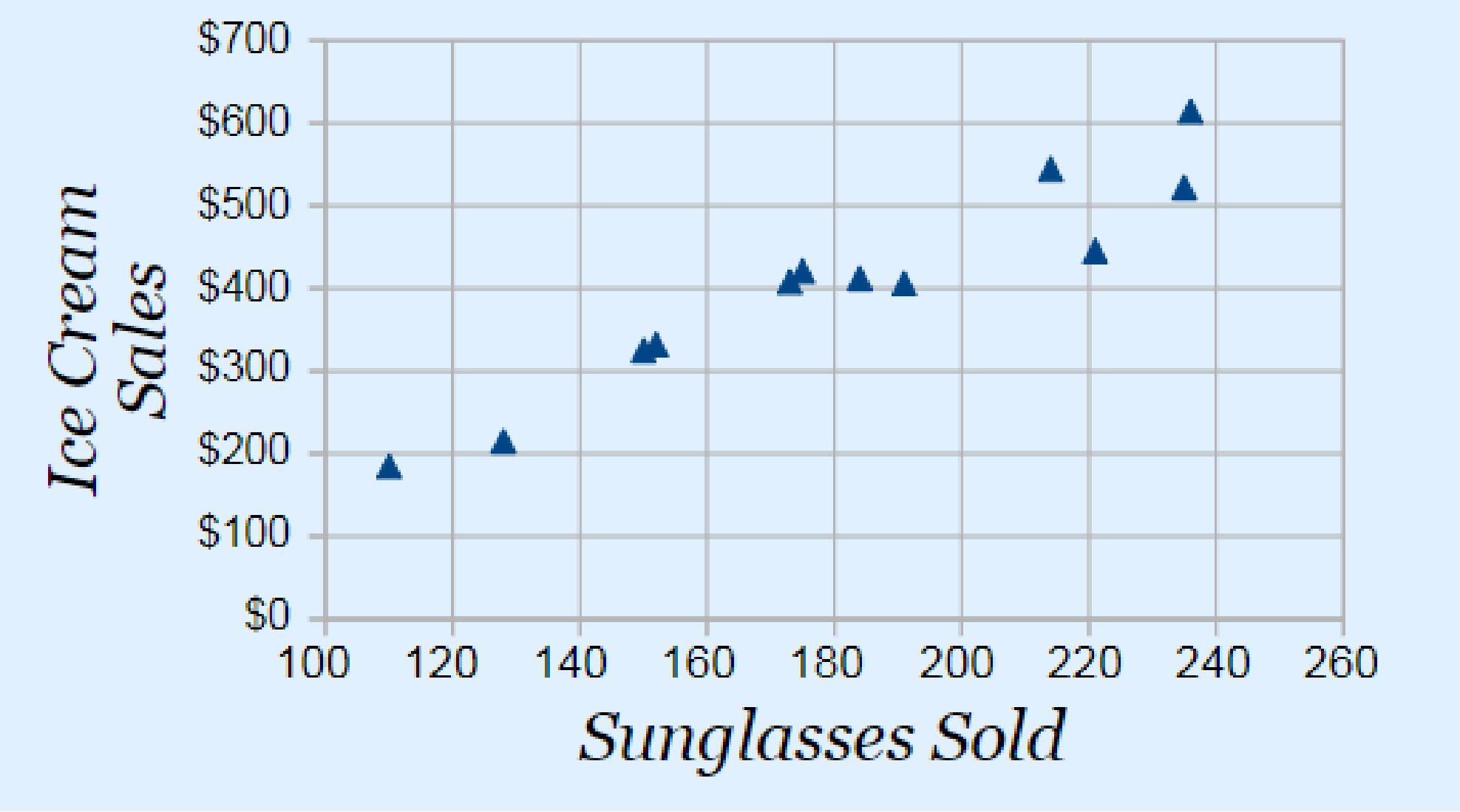


Correlation never proves causation




Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Correlation never proves causation



Multivariate Data Analysis

Identify pairs of multiple variables which could be interesting

Visualise

Slice and drill-down

Explain!

What charts could we use?

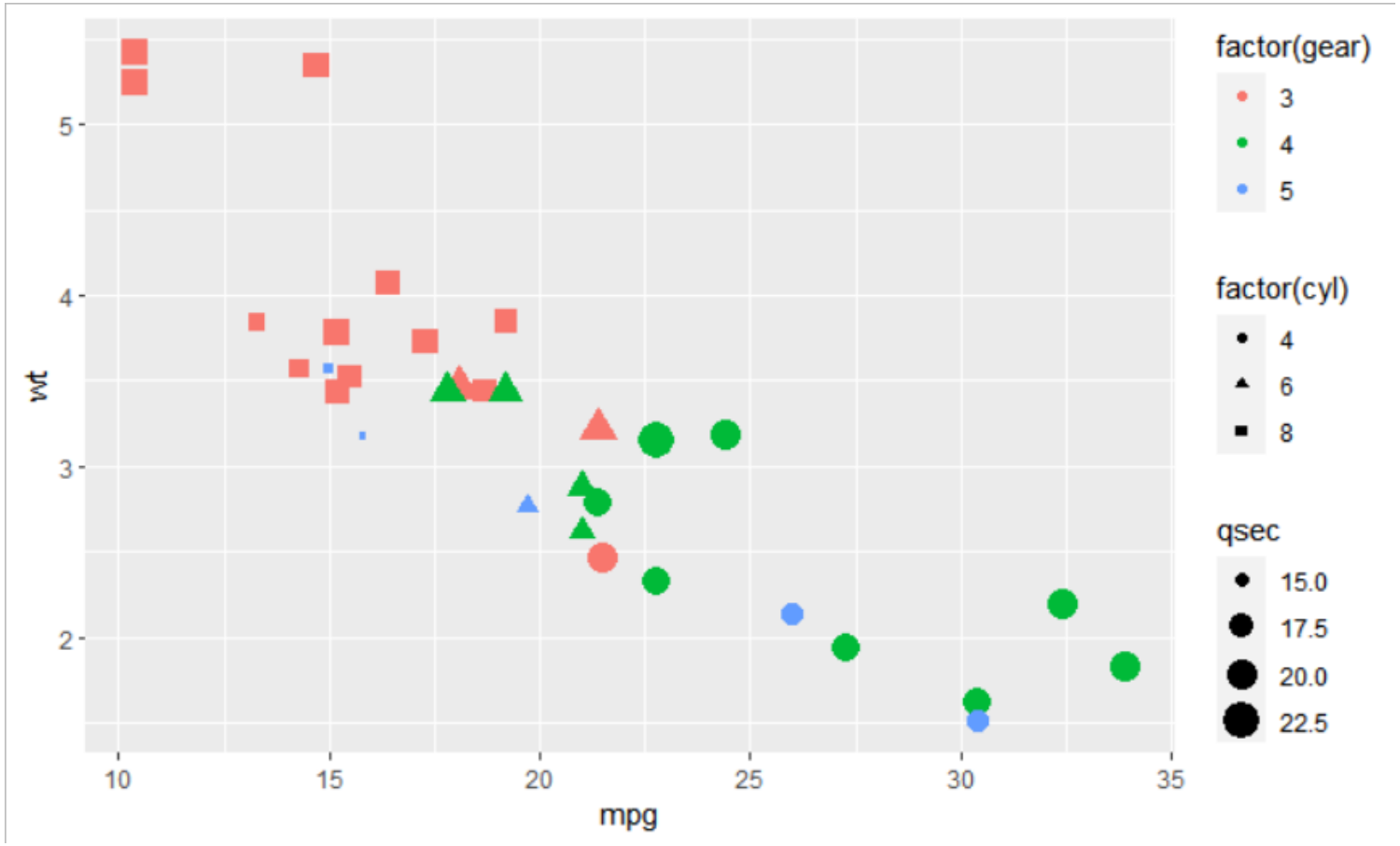
Scatter plot

How many dimensions can we plot in a scatter plot?

Continuous variables: x, y, size

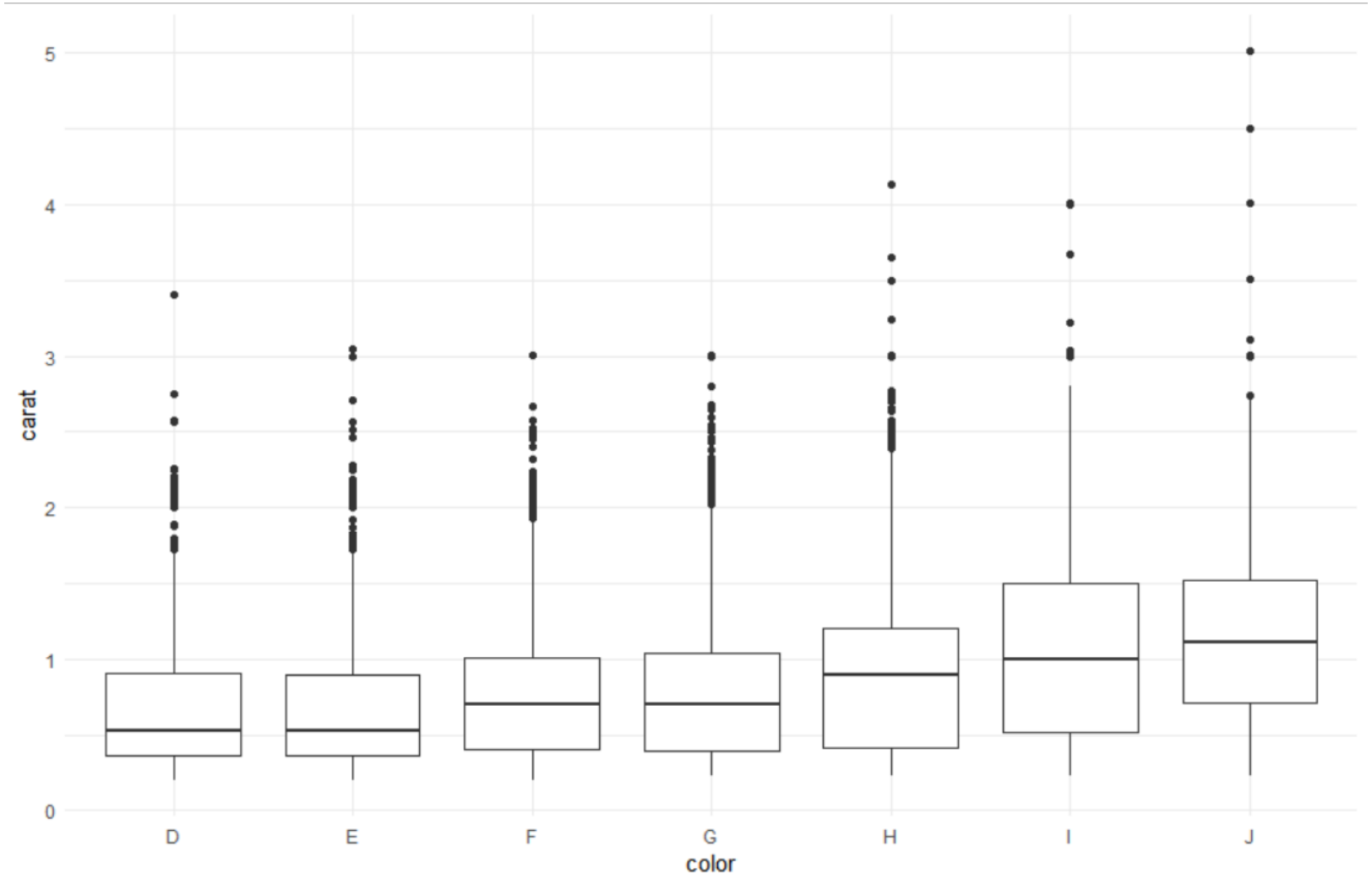
Discrete variables: colour, shape

Take care not to plot too many variables, generally two continuous and one discrete variable



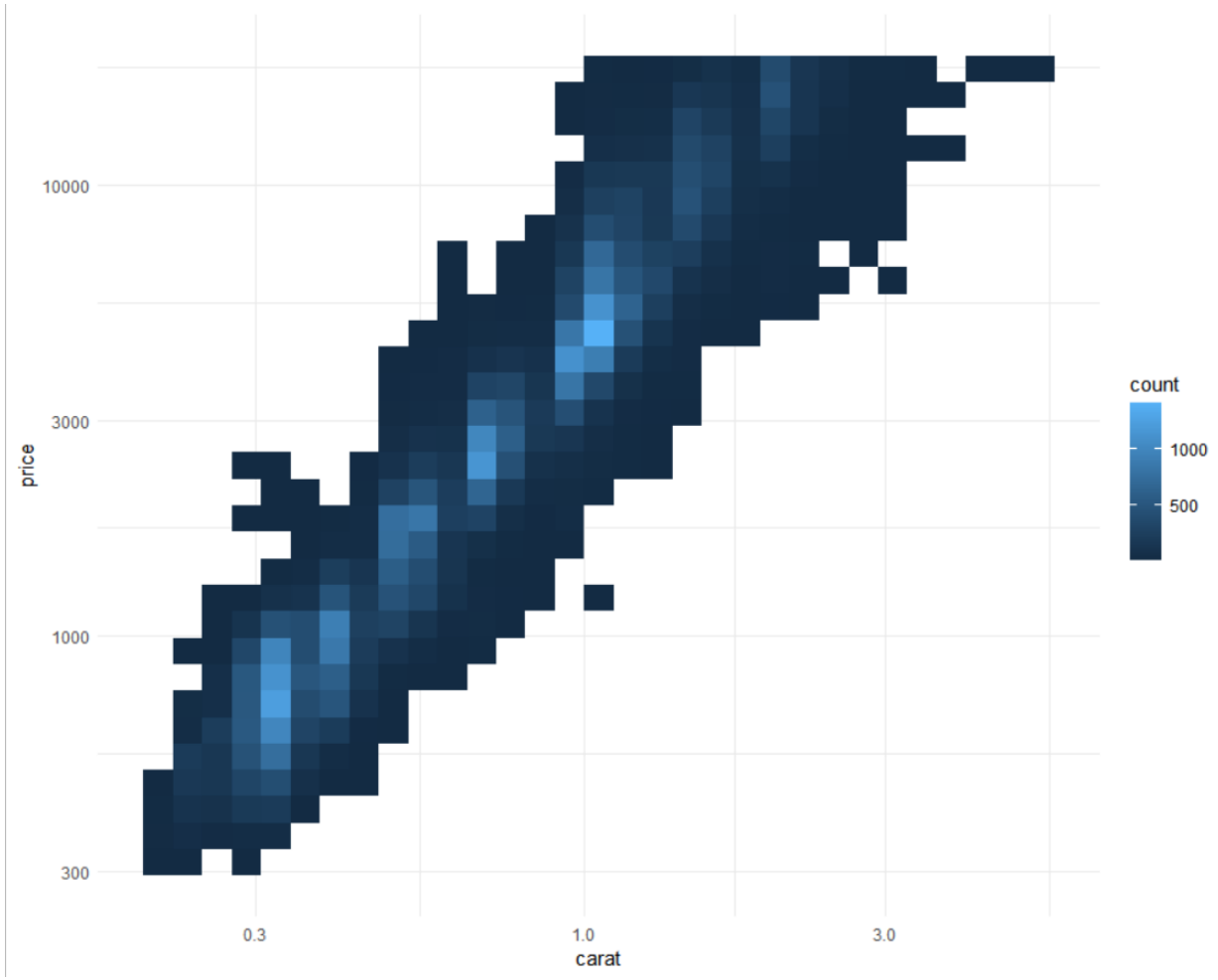
Box plot (again)

Show distribution of continuous variable by category



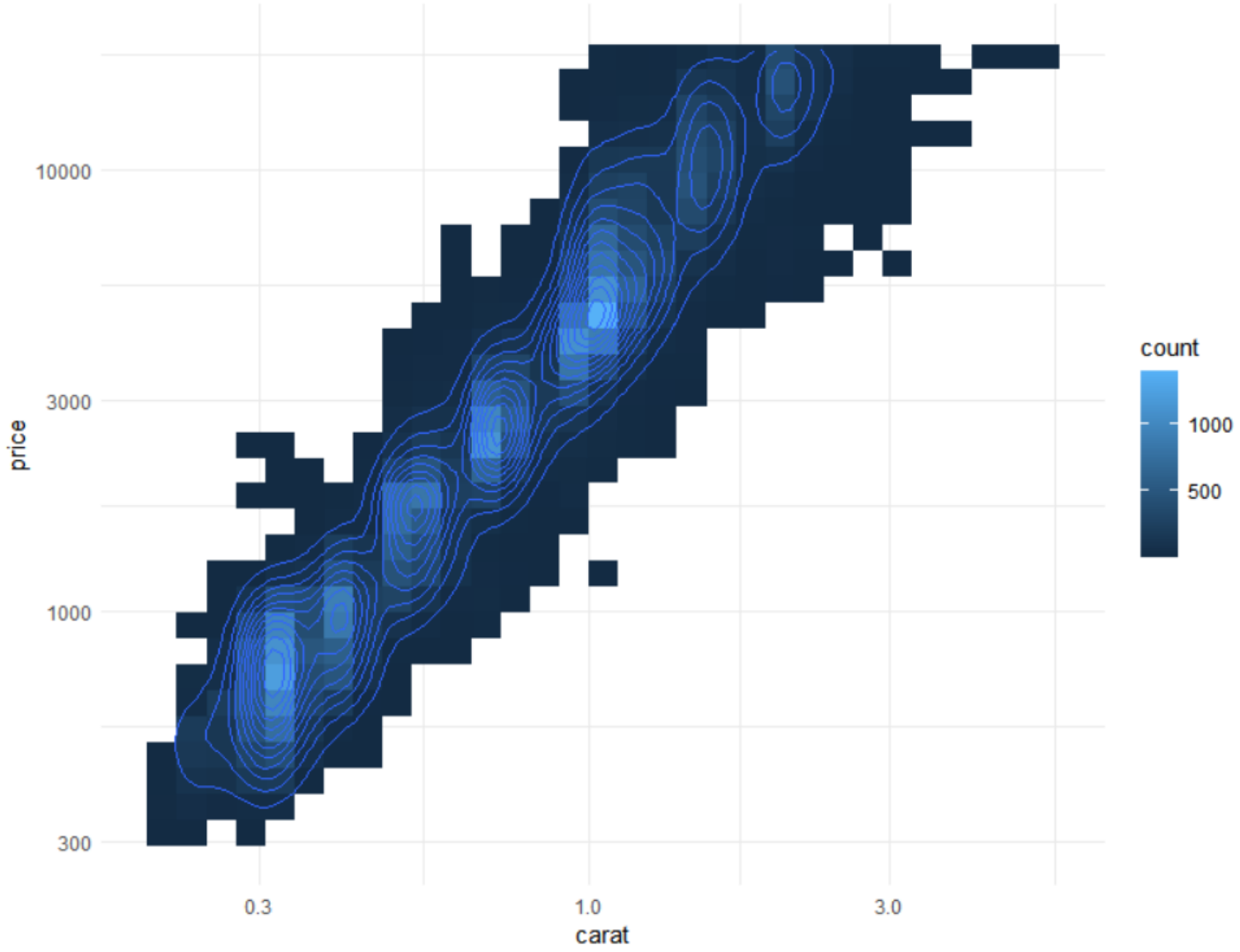
Heatmap

Shows the distribution of two continuous variables



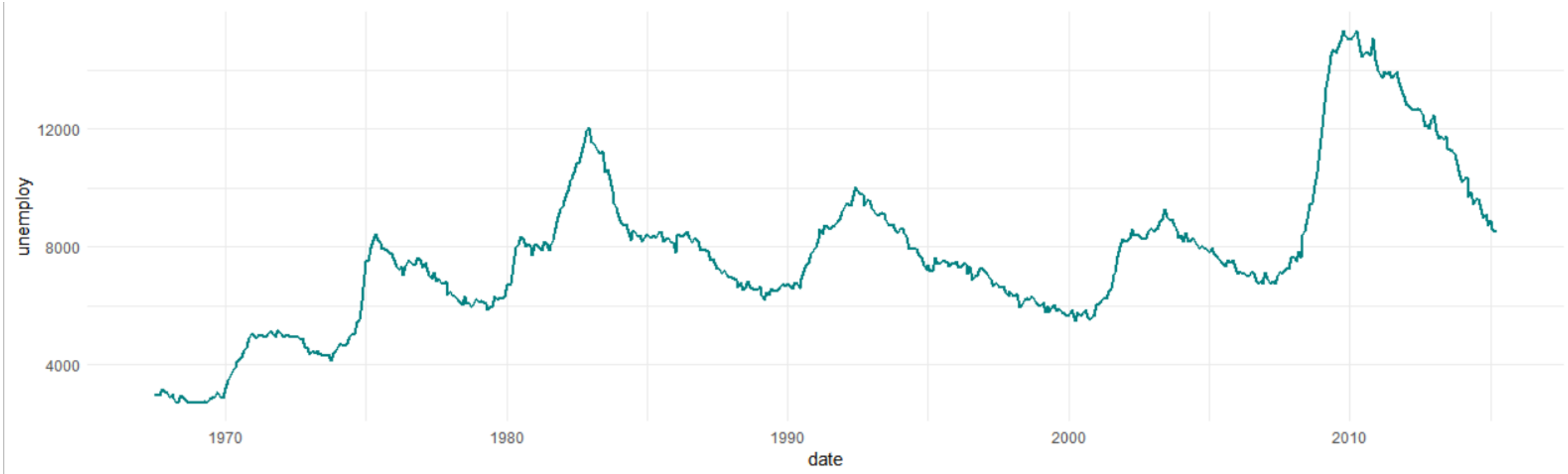
Two-dimensional density plot

Heat map with density map overlaid



Line chart

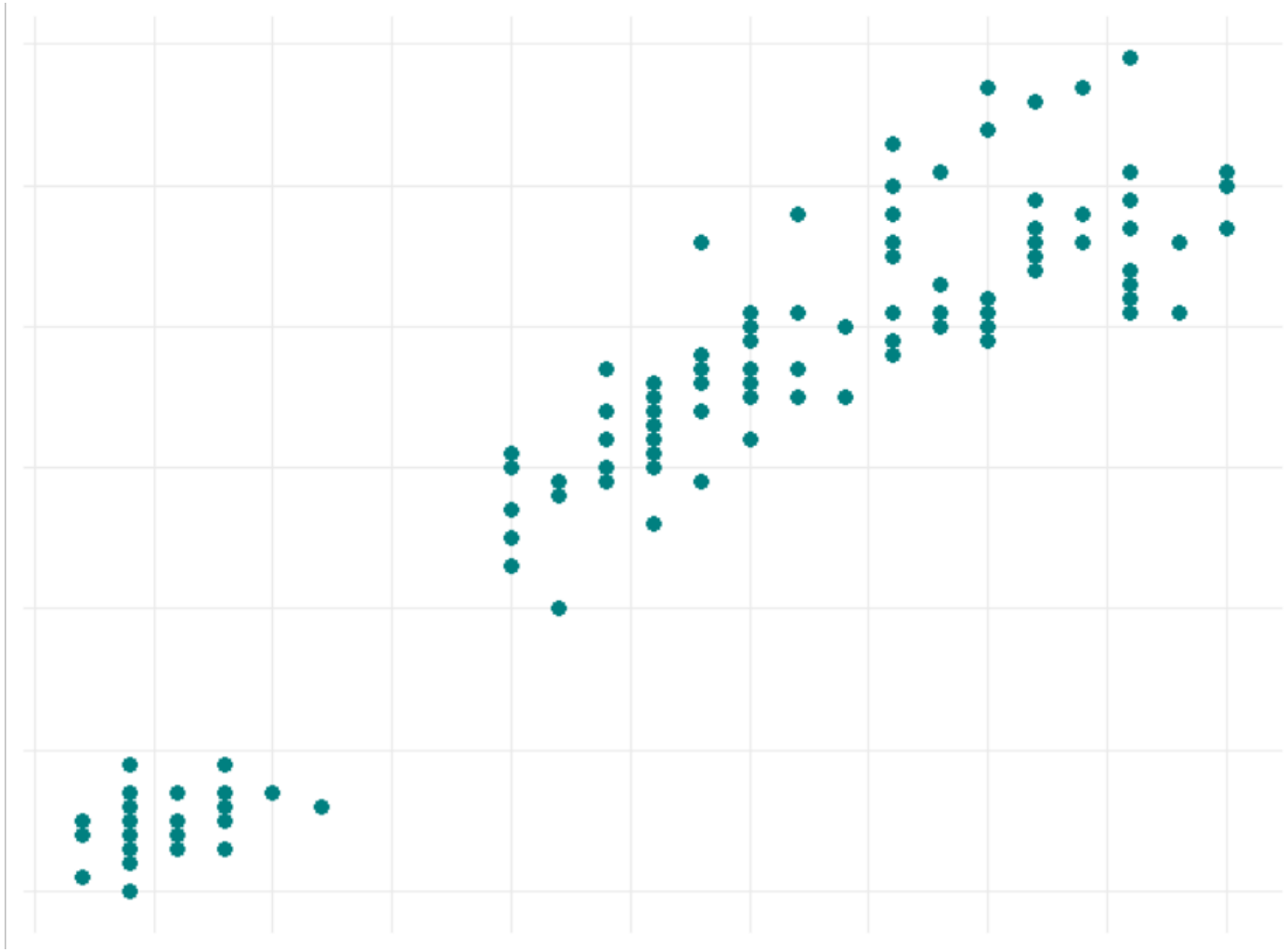
Continuous function (e.g. time series)



Grouping data

Segmenting data can expose behaviour which is not obvious in complete data set.

Smaller slices can be more manageable.

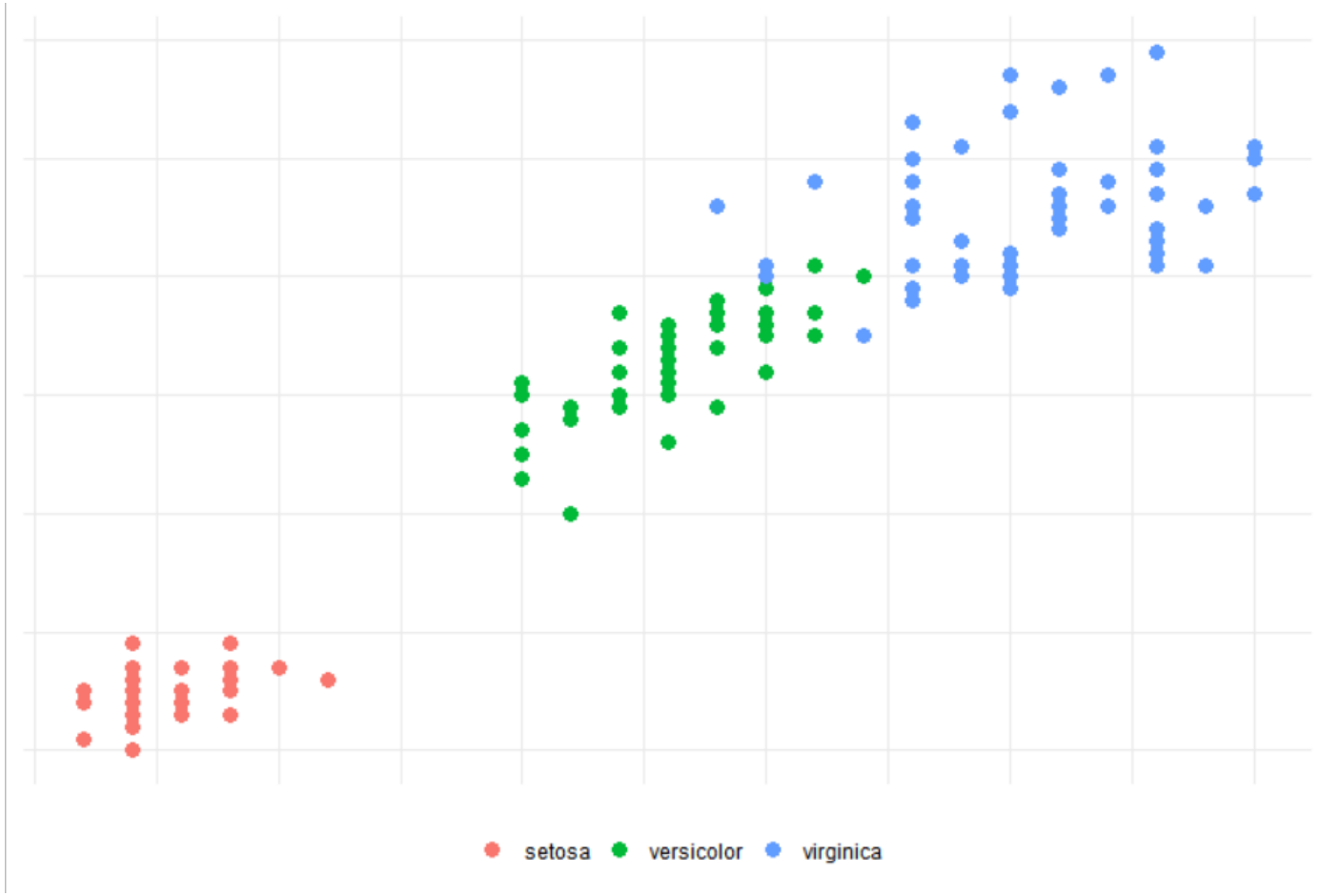


How to group?

Identify groups visually

Check if groups correlate to category

Try clustering to find a new dimension



Possible EDA strategies

- Start with basic, univariate questions
  - Frequency distributions?
  - Outliers?
- Move to multivariate plotting
  - Find relationships
- Segment data, if necessary and possible
- Finally, think about dimensionality reduction (e.g. PCA)

Remember…

EDA is a general approach – not prescriptive!

Keep asking questions of the data

Why are you observing a behaviour?

Once you find a plausible model, focus on residuals

data = fit + residuals

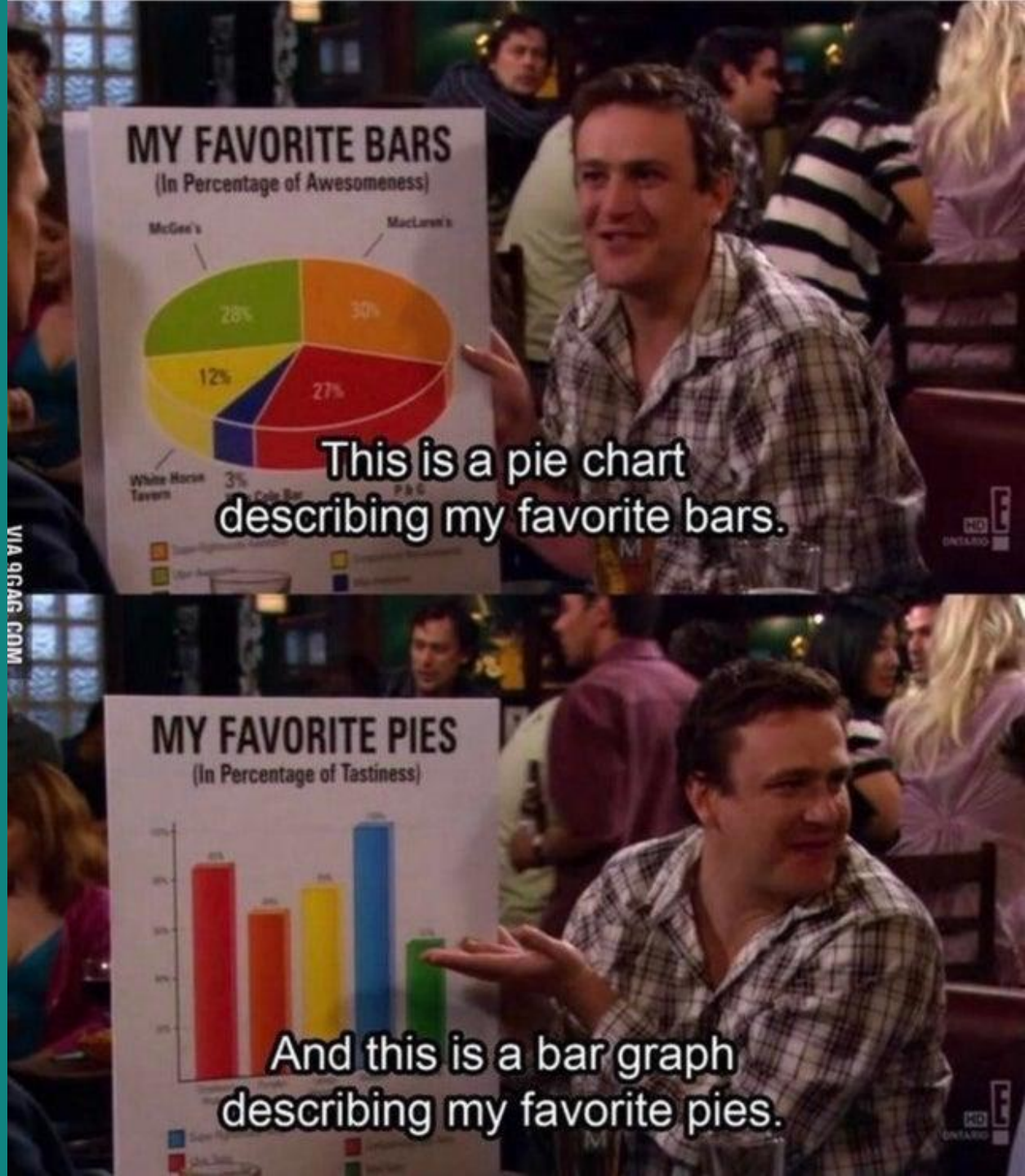
“Exploratory data analysis is detective work … (it) can never be the whole story but nothing else can serve as the foundation stone.”

— John Tukey

Chart of the week

—

Parallel coordinates chart



How many variables can you fit in one chart?

UNIVERSITY OF WESTMINSTER

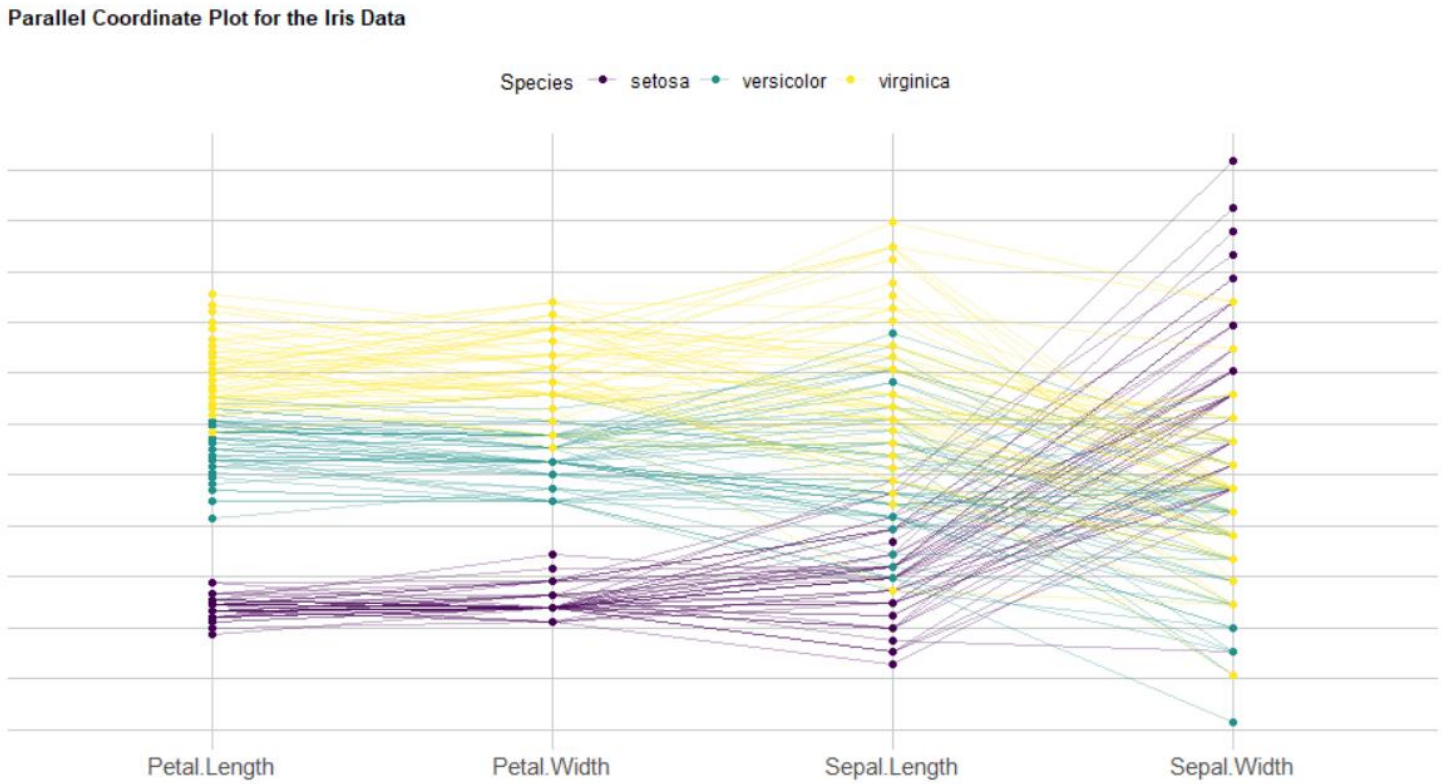
Parallel Coordinate Plot

Can display a large number of quantitative variables

Each observation is represented by a line

Colour can be used to differentiate categories

Collective “shape” helps to find pattern

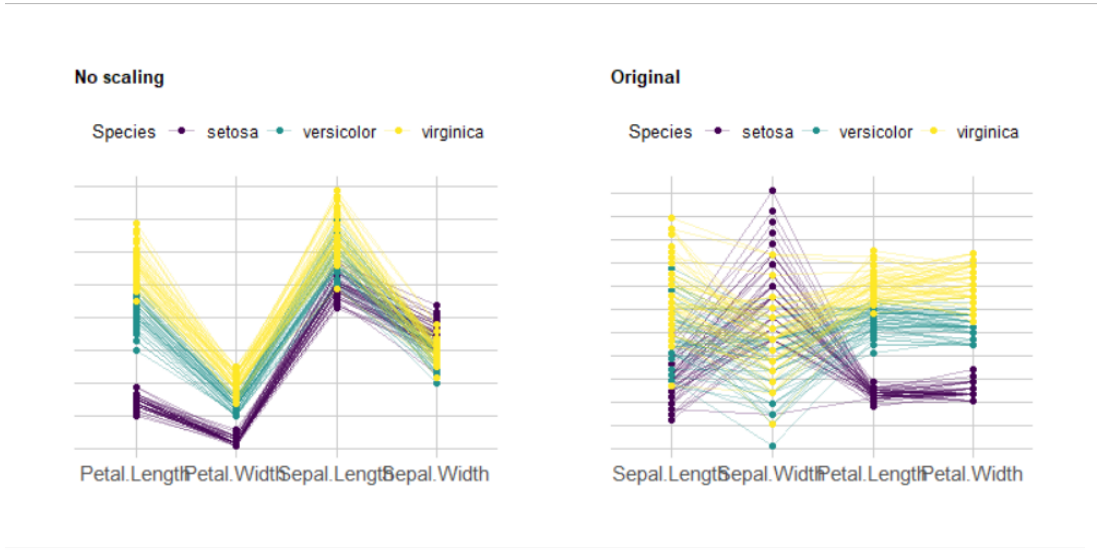


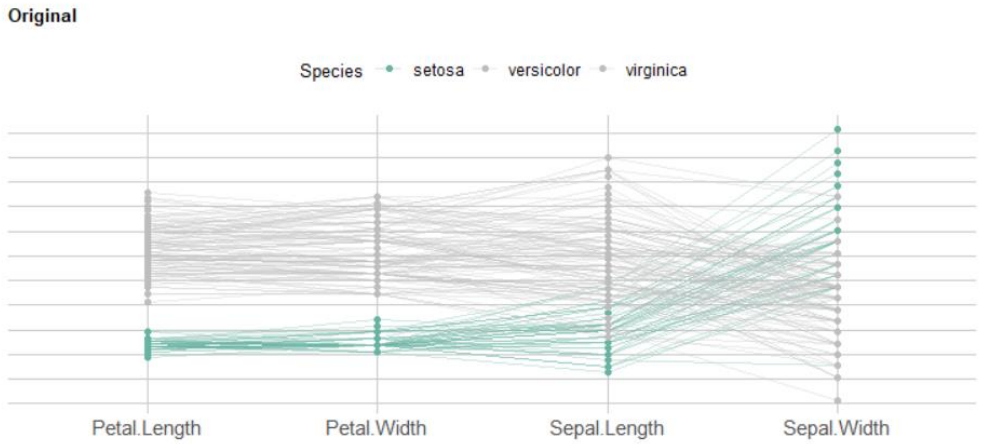
A few points to consider…

Scale all coordinates to the same scale

Be mindful of axis ordering

Highlight **interesting** category





UNIVERSITY OF
WESTMINSTER