What can we expect from the data?

Statistics, Visualisation, Insights and Fallacies

Jianqi Yang

Westlake University

Westlake Institute for Advanced Study

Table of contents I

Descriptive statistics and visualisation

Correlation & causation

Causal Inference and Machine Learning

Descriptive statistics and visualisation

Descriptive statistics

Descriptive statistics refers to a branch of statistics that involves summarizing, organizing, and presenting data meaningfully and concisely. It focuses on describing and analyzing a dataset's main features and characteristics without making any generalizations or inferences to a larger population.

Example 1

How do you describe the data below?

```
      x    y
1  55.4 97.2
2  51.5 96.0
3  46.2 94.5
4  42.8 91.4
5  40.8 88.3
6  38.7 84.9
7  35.6 79.9
8  33.1 77.6
9  29.0 74.5
10 26.2 71.4
```

`and so many`

Example 1

Naturally, we will use descriptive statistics:

$mean_x$	var_x	$mean_y$	var_y	cor
39.93	88.49	85.57	87.19	0.98

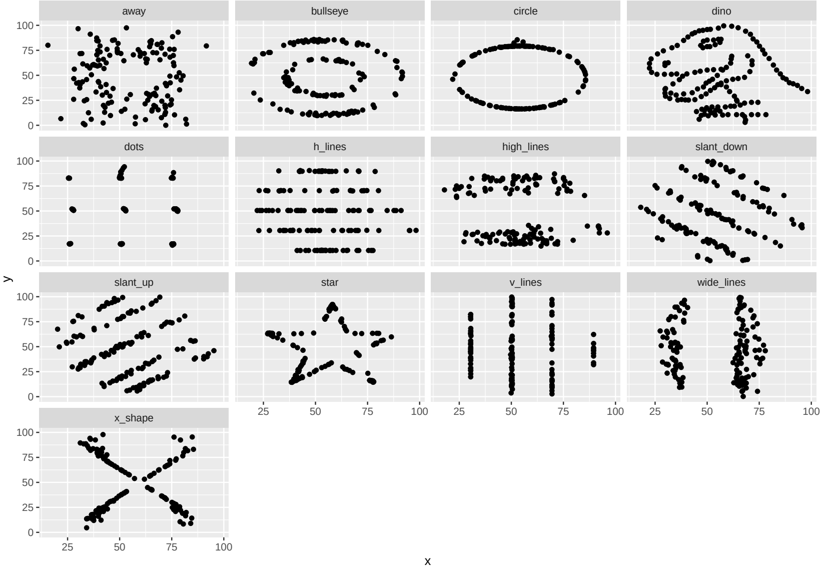
Example 1

So what might this dataset be?

- Some common points
- A dinosaur
- A star
- Scratches left by kittens
- X-Wing (May the Force be with you!)

Example 1

Hopefully the X-Wing won't ruin the composition...



Example 1

dataset	mean_x	var_x	mean_y	var_y	cor
away	54.27	281.23	47.83	725.75	-0.06
bullseye	54.27	281.21	47.83	725.53	-0.07
circle	54.27	280.90	47.84	725.23	-0.07
dino	54.26	281.07	47.83	725.52	-0.06
dots	54.26	281.16	47.84	725.24	-0.06
h_lines	54.26	281.10	47.83	725.76	-0.06
high_lines	54.27	281.12	47.84	725.76	-0.07
slant_down	54.27	281.12	47.84	725.55	-0.07
slant_up	54.27	281.19	47.83	725.69	-0.07
star	54.27	281.20	47.84	725.24	-0.06
v_lines	54.27	281.23	47.84	725.64	-0.07
wide_lines	54.27	281.23	47.83	725.65	-0.07
x_shape	54.26	281.23	47.84	725.22	-0.07

Regression Analysis

Similar phenomena exist in more advanced forms of descriptive statistics, such as regression.

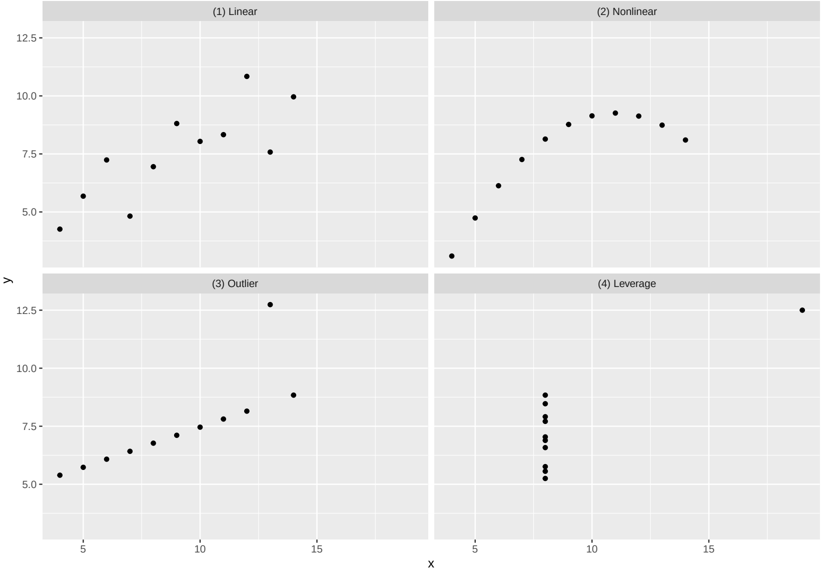
Example 2

Based on Example 1, we have learnt that descriptive statistics are not sufficient to properly represent the nature of the dataset and that visualisation using scatterplots is a natural choice.

Consider the following dataset:

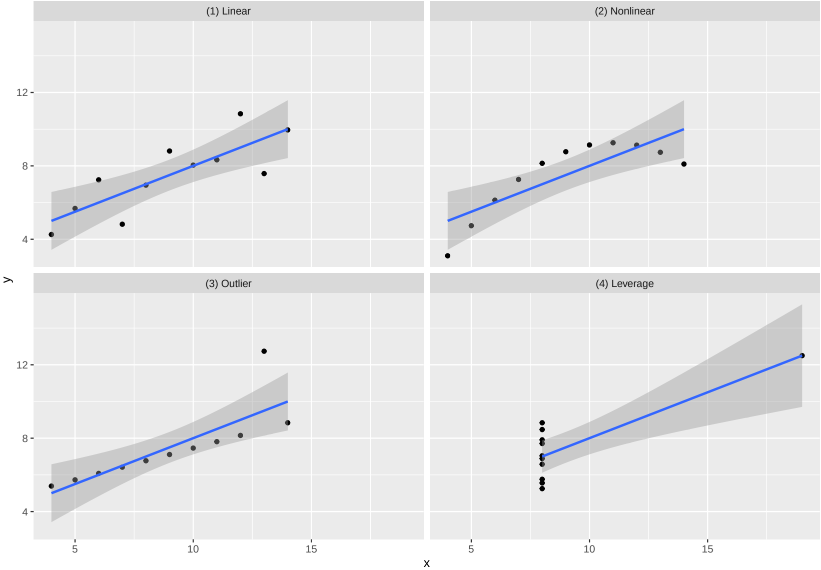
- mean of x: 9
- variance of x: 11
- mean of y: 7.5
- variance of y: 4.125
- correlation between x and y: 0.816

Example 2

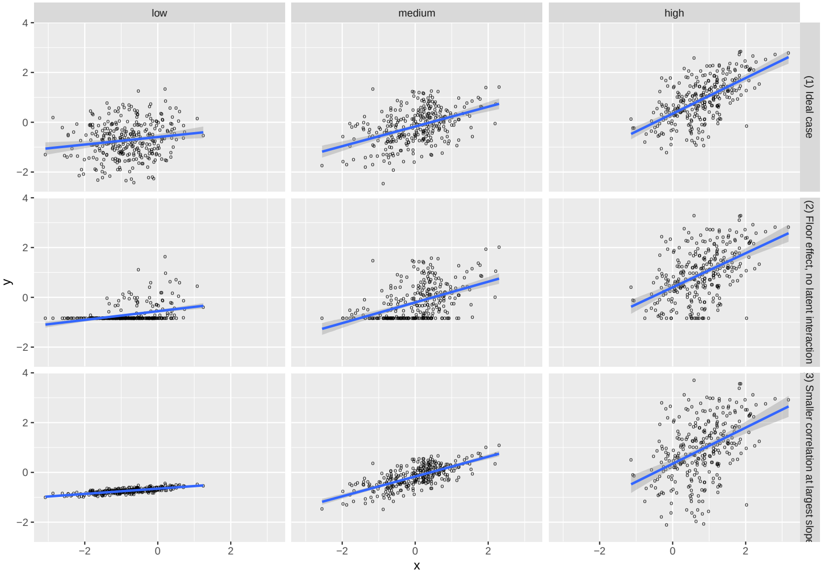

(1) Linear
(2) Nonlinear
(3) Outlier
(4) Leverage
y
x
12.5
10.0
7.5
5.0
12.5
10.0
7.5
5.0
5
10
15
5
10
15

Example 2

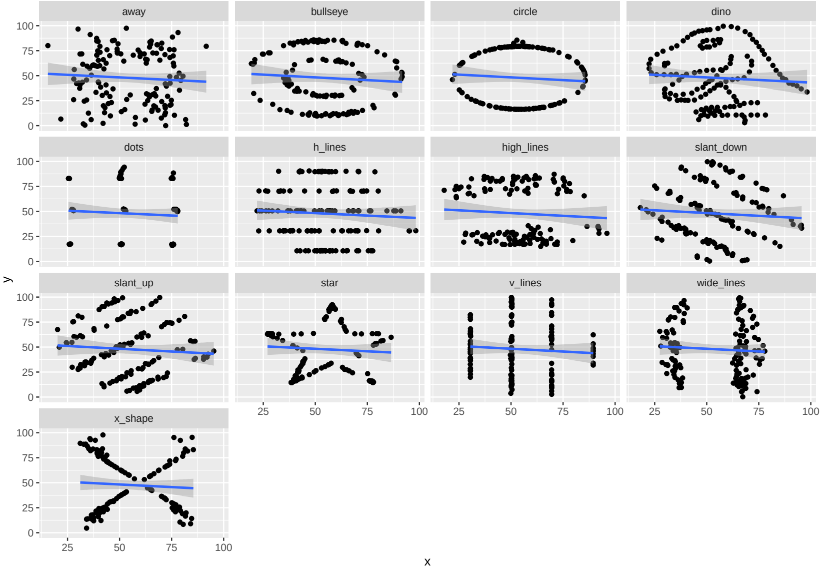
- $y = 0.5x + 3,\ R^2$: 0.67



Example 3


low
medium
high
(1) Ideal case
(2) Floor effect, no latent interaction
3) Smaller correlation at largest slope
y
4
2
0
−2
x
−2
0
2

Example 4


away
bullseye
circle
dino
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines
x_shape
y
x
0
25
50
75
100
25
50
75
100

Rethinking

- Descriptive statistics are not actually describing
- Statistical methods always entail a loss of information
- Don't trust methods such as regression and, in particular, don't confuse correlation with causation

I will cover the third point in the next section

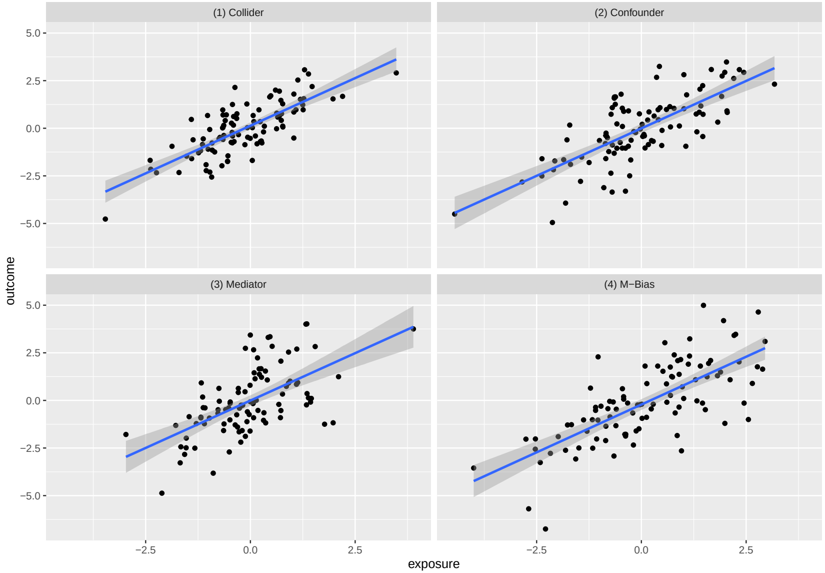
Correlation & causation

An argument

Which is correct?

- correlation proves causation
- correlation does not imply causation

Example 5


(1) Collider
(2) Confounder
(3) Mediator
(4) M−Bias
outcome
exposure
5.0
2.5
0.0
−2.5
−5.0
−2.5
0.0
2.5

Example 5

Data generating mechanism	Y ~ X	Y ~ X + Z	Corr of X and Z
(1) Collider	1	0.55	0.7
(2) Confounder	1	0.50	0.7
(3) Mediator	1	0.00	0.7
(4) M-Bias	1	0.88	0.7

Example 5

Collider:

$$
\begin{aligned}
&X \sim N(0,1)\\
&Y = X + \varepsilon_y, \varepsilon_y \sim N(0,1)\\
&Z = 0.45X + 0.77Y + \varepsilon_z, \varepsilon_z \sim N(0,1)
\end{aligned}
$$

Example 5

Confounder:

$$
\begin{aligned}
&Z \sim N(0,1) \\
&X = Z + \varepsilon_x, \varepsilon_x \sim N(0,1) \\
&Y = 0.5X + Z + \varepsilon_y, \varepsilon_y \sim N(0,1)
\end{aligned}
$$

Example 5

Mediator:

$$\begin{aligned}
&X \sim N(0,1)\\
&Z = X + \varepsilon_z, \varepsilon_z \sim N(0,1)\\
&Y = Z + \varepsilon_y, \varepsilon_y \sim N(0,1)
\end{aligned}$$

Example 5

M-Bias:

$$
\begin{aligned}
&U_1 \sim N(0,1)\\
&U_2 \sim N(0,1)\\
&Z = 8U_1 + U_2 + \varepsilon_z, \varepsilon_z \sim N(0,1)\\
&X = U_1 + \varepsilon_x, \varepsilon_x \sim N(0,1)\\
&Y = X + U_2 + \varepsilon_y, \varepsilon_y \sim N(0,1)
\end{aligned}
$$

Example 6

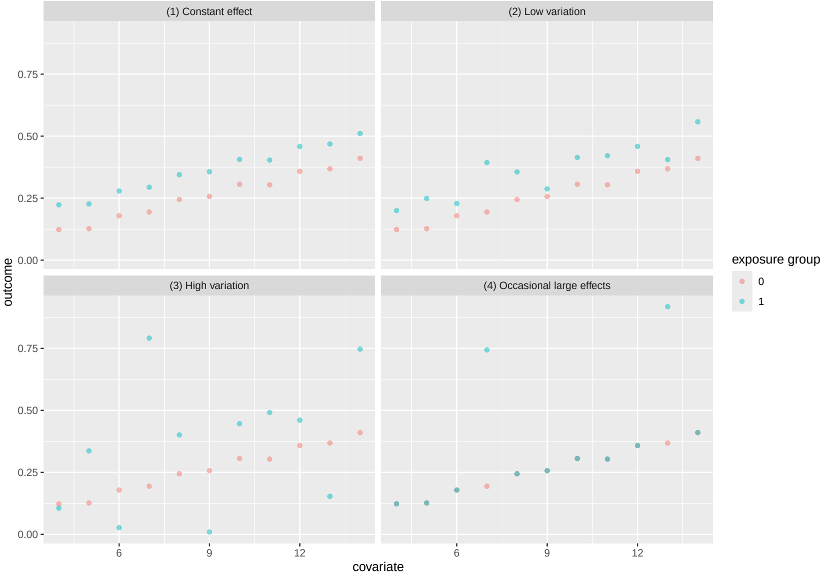
ATE: Average Treatment Effect, a measure used to compare treatments (or interventions) in randomized experiments, evaluation of policy interventions, and medical trials.

Example 6.1

We can get the same average treatment effect despite variability across some pre-treatment characteristic (here called covariate).

dataset	ATE
(1) Constant effect	0.1
(2) Low variation	0.1
(3) High variation	0.1
(4) Occasional large effects	0.1

Example 6.1

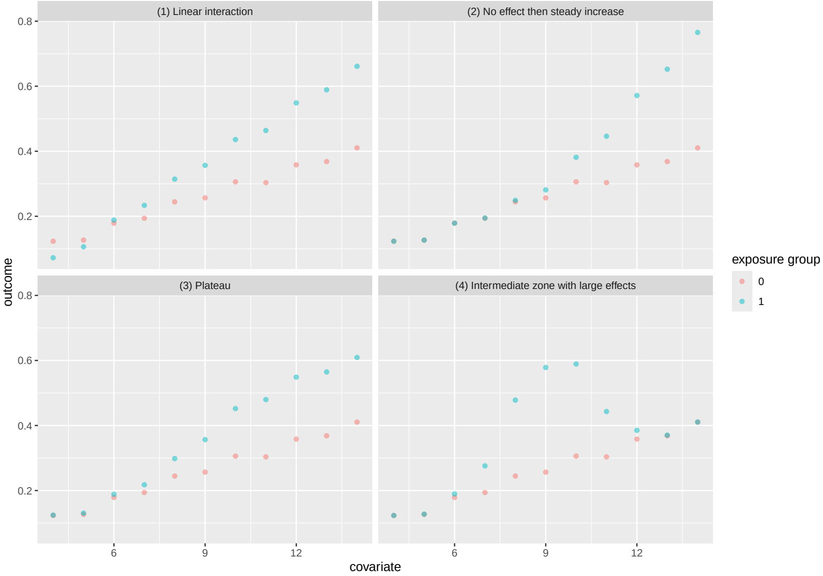

(1) Constant effect
(2) Low variation
(3) High variation
(4) Occasional large effects
outcome
0.00
0.25
0.50
0.75
covariate
6
9
12
exposure group
● 0
● 1

Example 6.2

We can observe the same causal effect under different patterns of treatment heterogeneity.

dataset	ATE
(1) Linear interaction	0.1
(2) No effect then steady increase	0.1
(3) Plateau	0.1
(4) Intermediate zone with large effects	0.1

Example 6.2


(1) Linear interaction
(2) No effect then steady increase
(3) Plateau
(4) Intermediate zone with large effects
outcome
0.8
0.6
0.4
0.2
covariate
6
9
12
exposure group
0
1

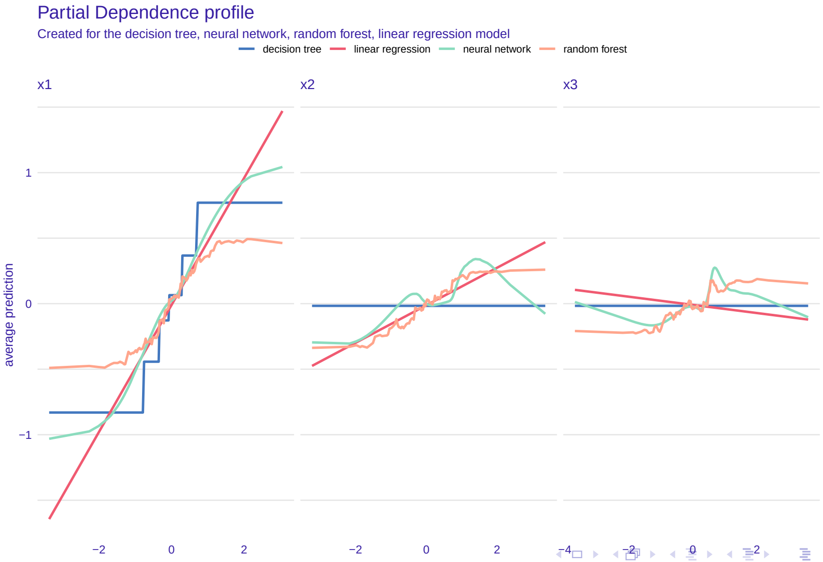
Causal Inference and Machine Learning

Introduction

This phenomenon exists not only in the analysis of data sets, but also in the analysis of models. In essence, all models from linear regression to neural networks are functions abstracted from the dataset.

Example 7

The three models clearly follow completely different design ideas and training processes, as can be easily noticed from the images:



Example 7

However, they all have the same R^2 and $RMSE$.

model	R2	RMSE
Decision tree	0.73	0.35
Linear regression	0.73	0.35
Random forest	0.73	0.35
Neural network	0.74	0.35

Conclusion

- Even the most sophisticated models that are data-driven only always face a Hume problem.
- It should be borne in mind that causal inferences are always ‘inferences’ and remain in fact causeless.
- Proper physical understanding and intuition of the system under study is the true foundation of data science.