UNIVERSITY OF ABERDEEN

CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

Relationships In Data

Data Mining & Visualisation
Lecture 5

2025

Recap...

- Categorical vs. Quantitative data
- Levels of measurement

Today...

- Relationships in data
- Pearson Correlation Coefficient

Understanding Relationships in Data

Often, there will be relationships between variables in data.

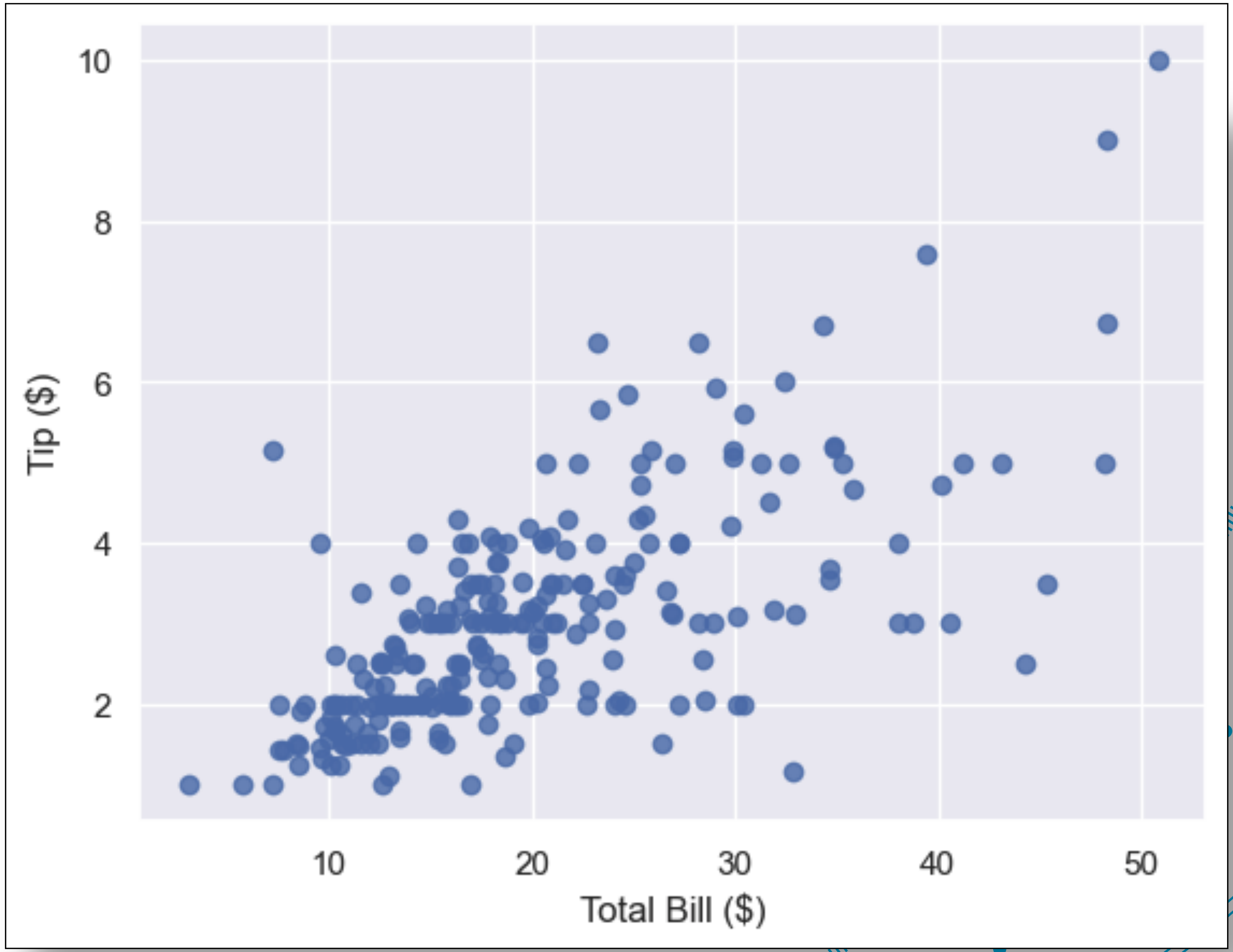
E.g. as X increases, what *typically* happens to the value of Y?

Data mining can help us to uncover and understand these relationships, and provide us with ways to visualise them.

Understanding Relationships in Data

Example: Do larger bills result in larger tips?

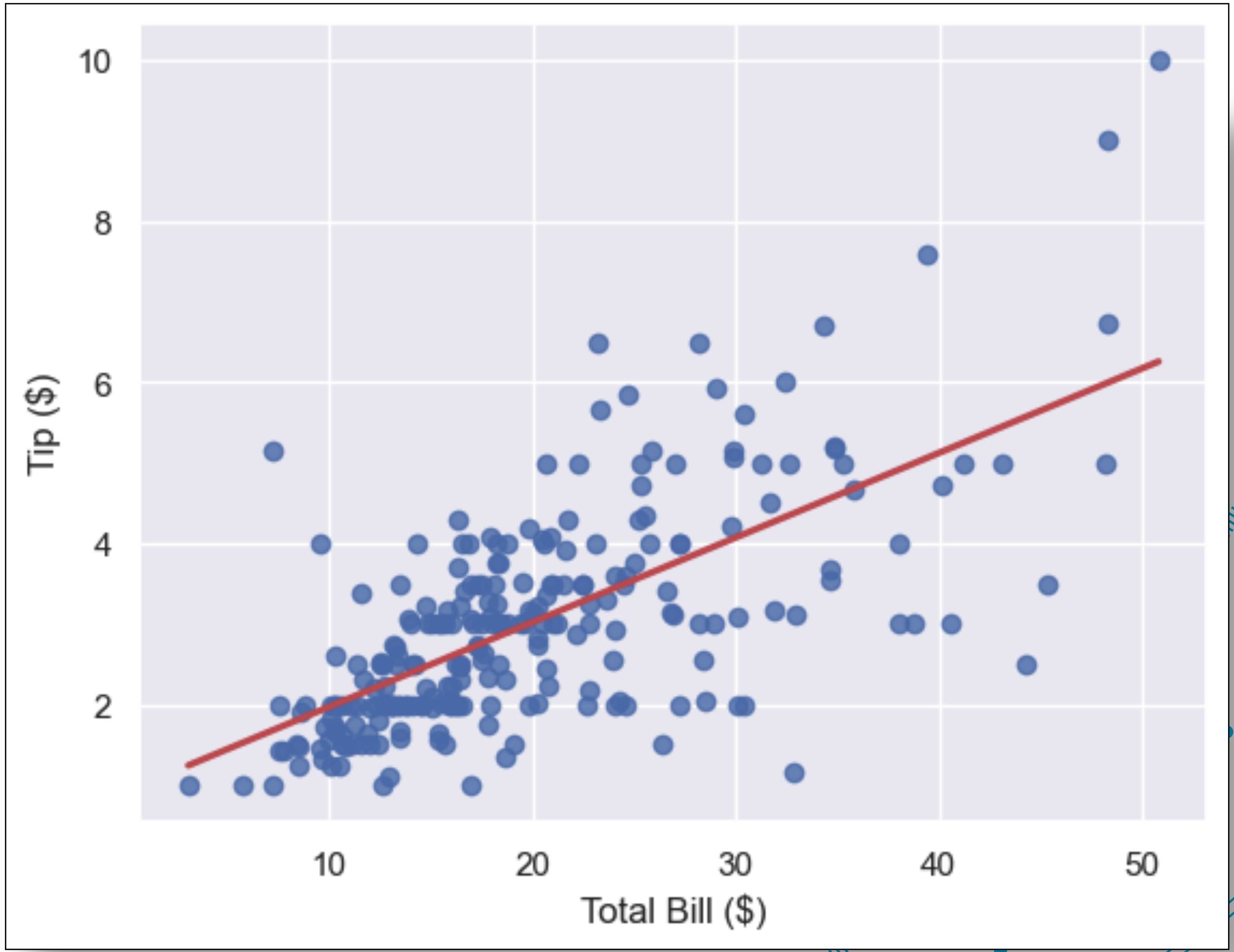
As the total bill increases, what *typically* happens to the value of tips?



Understanding Relationships in Data

Example: Do larger bills result in larger tips?

As the total bill increases, what *typically* happens to the value of tips?



Correlation

In statistics, 'correlation' is a measurement of the strength of relationship between two variables.

When people talk about a correlation coefficient, they're often referring to the Pearson Correlation Coefficient (PCC).

Strength and Direction

It is a numerical value between –1 and 1, representing the strength of association.

Correlations can be described as positive (>0) or negative (<0).

They can also be described as strong or weak, referring to the strength of the correlation (either positive or negative).

Strength and Direction

A correlation of **0** indicates no correlation between the variables.

As the correlation coefficient reaches **1**, it represents the strongest possible correlation.

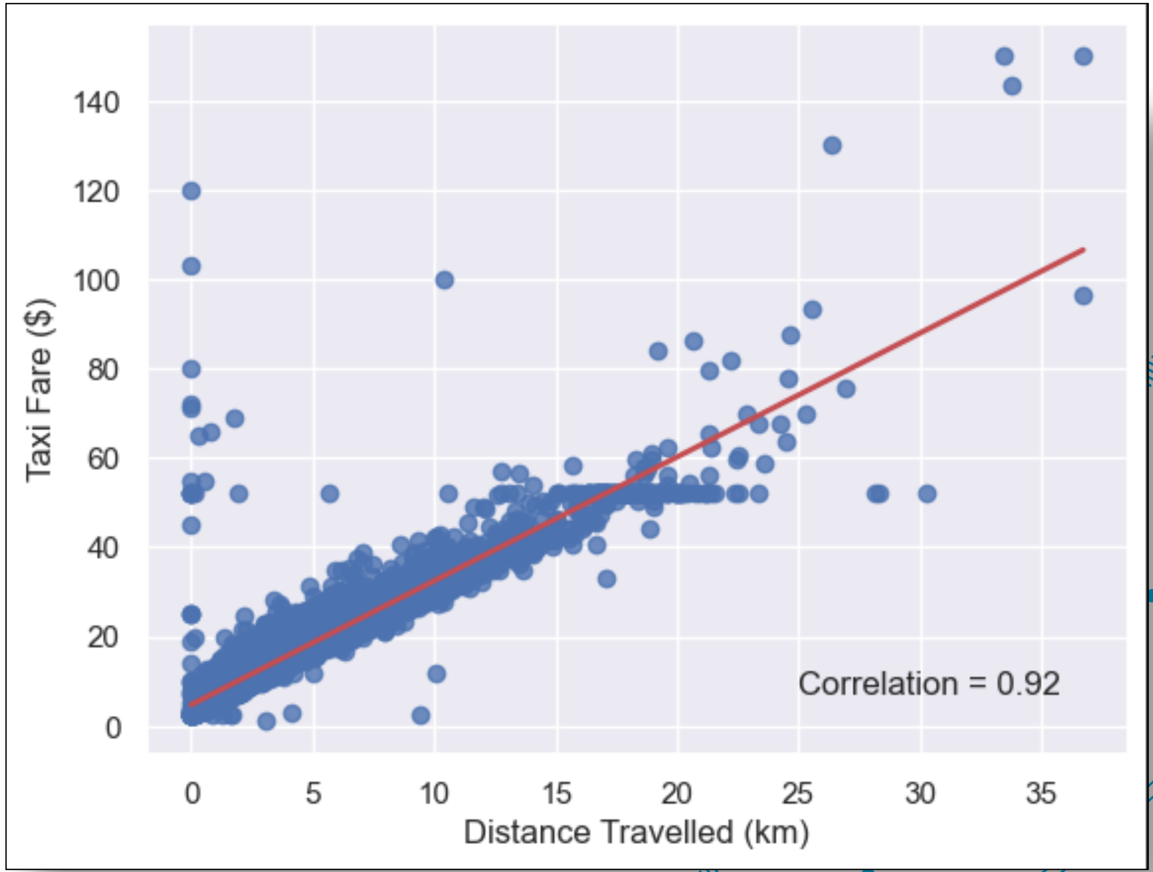
A correlation coefficient of -1 represents an inverse relationship.

Strong Positive Correlation

Let's look at the Taxis dataset.

Is there a correlation between a taxi's fare and the distance travelled?

There is a strong positive correlation (p=.92) between these two variables.

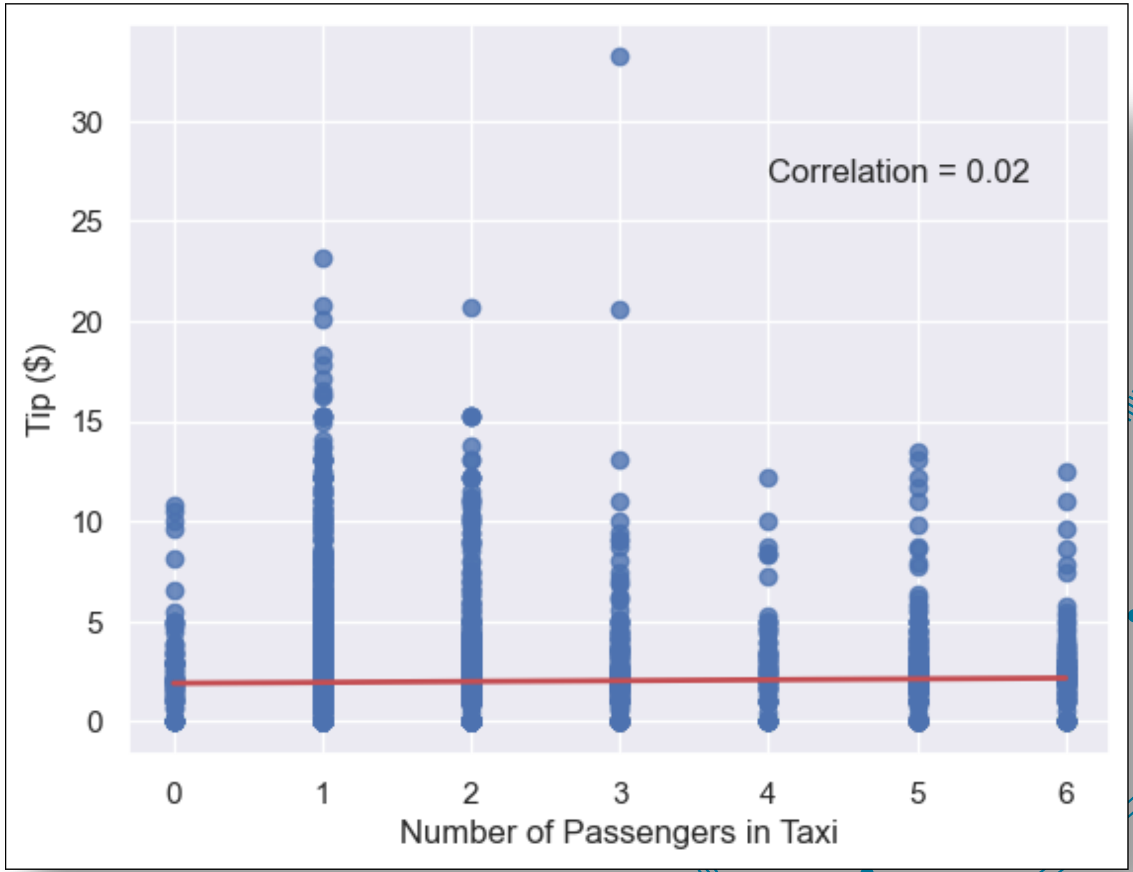


Weak Correlation

Let's look at the Taxis dataset.

Is there a correlation between the number of passengers and the amount tipped?

There is a very weak correlation (p=.02) between these two variables.

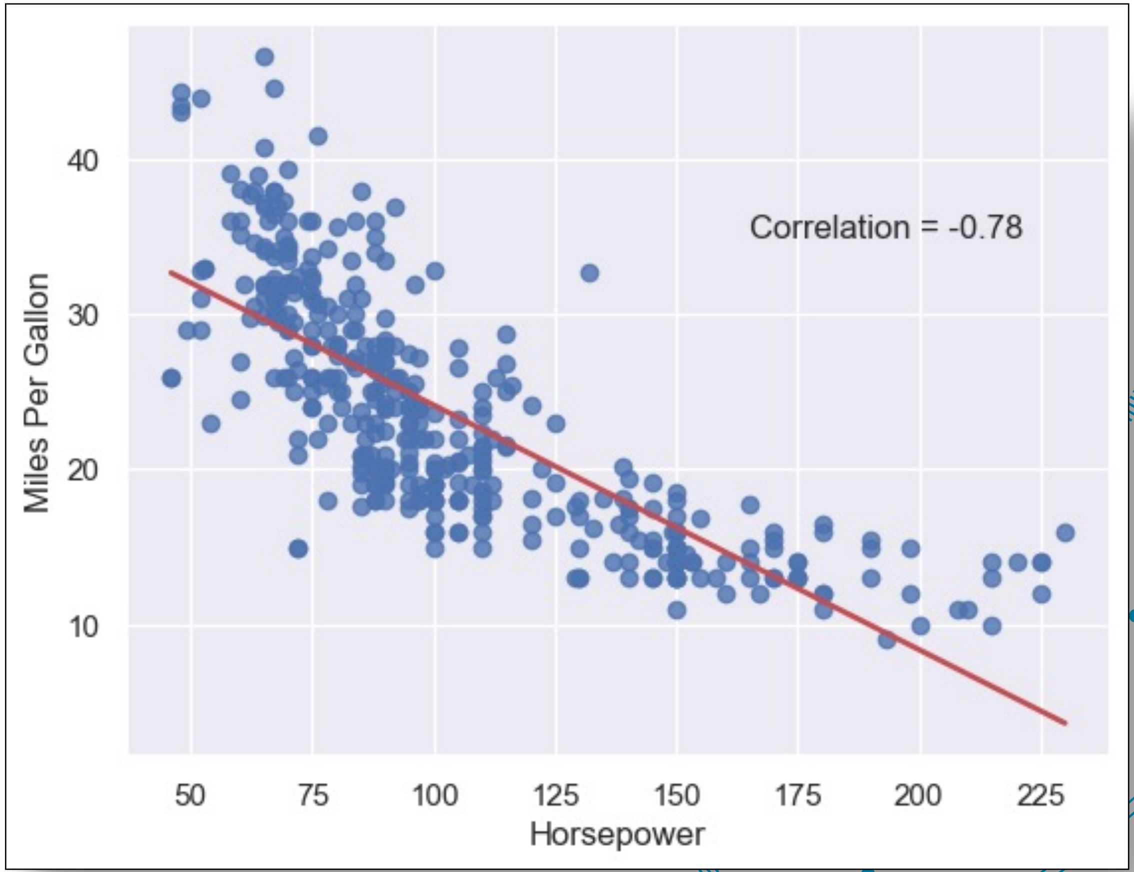


Negative Correlation

Let's look at the MPG dataset.

Is there a correlation between a car's horsepower and its miles per gallon?

There is a reasonably strong negative correlation (p = -.78) between these two variables.



Calculating the Pearson Correlation Coefficient of Two Variables

Note that there is a Pearson Correlation Coefficient for both **populations** and **samples**.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

(Population correlation coefficient)

There are slight differences in how you calculate these, but for simplicity, we will just focus on samples.

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{s_X s_Y}$$

(Sample correlation coefficient)

Calculating the Pearson Correlation Coefficient of Two Variables (Sample)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{14}{\sqrt{14*14}} = \frac{14}{14} = 1$$

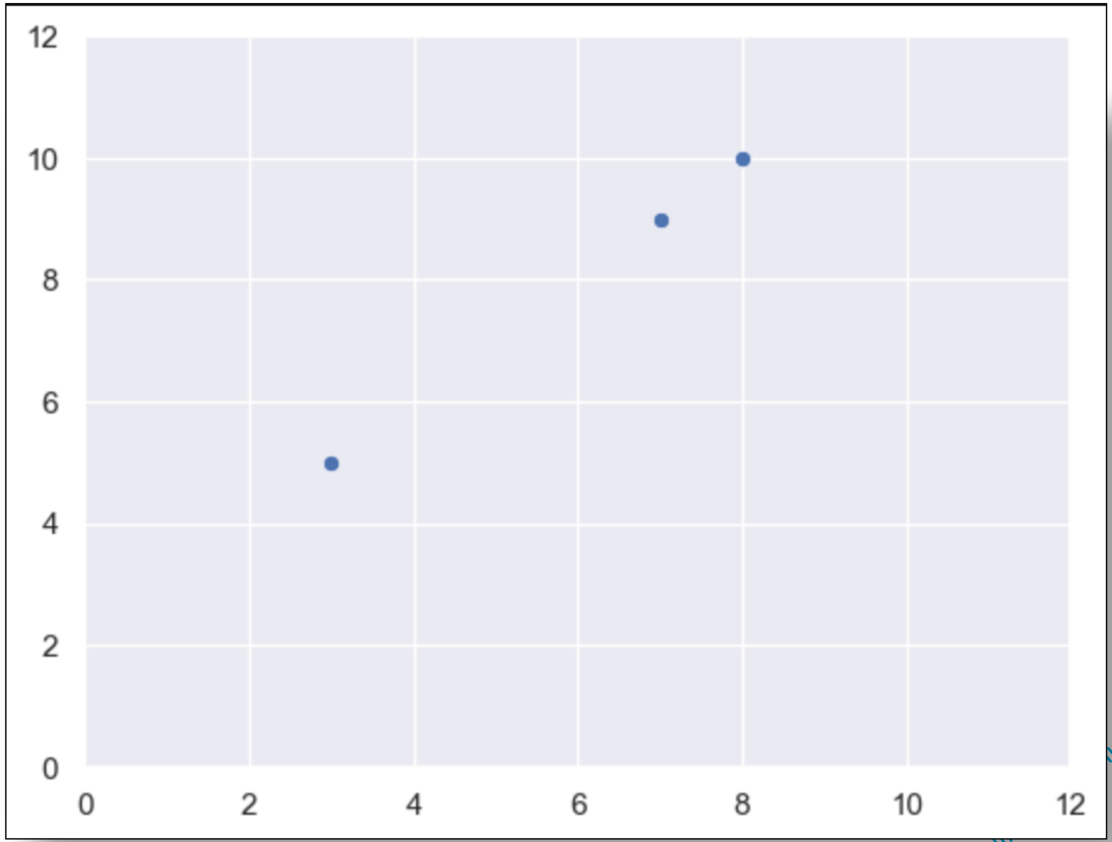
$\bar{x} = 6$

$\bar{y} = 8$

	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	3	5	-3	-3	9	9	9
	7	9	1	1	1	1	1
	8	10	2	2	4	4	4
$\sum$	**18**	**24**			**14**	**14**	**14**

Calculating the Pearson Correlation Coefficient of Two Variables (Sample)

x	y
3	5
7	9
8	10



Limitations of Correlation

Limitations of Correlation

You may have heard the phrase *'correlation does not imply causation'*.

But what does that actually mean?

Let's look at the definitions of these…

Limitations of Correlation

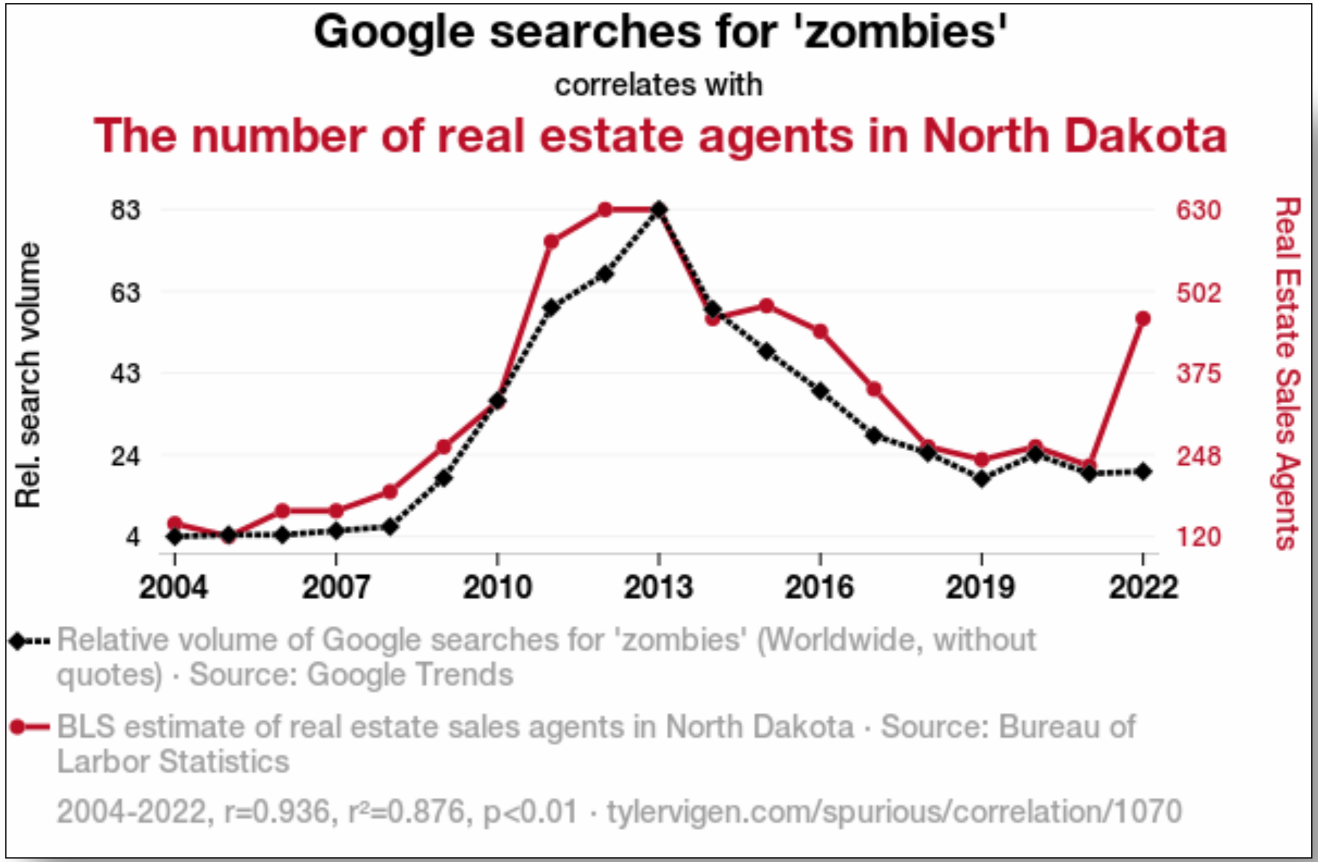
Correlation is a measure of the strength of an association between two variables.

- As variable x changes, to what extent does variable y change?

Causation is the process of one event causing or producing another event.

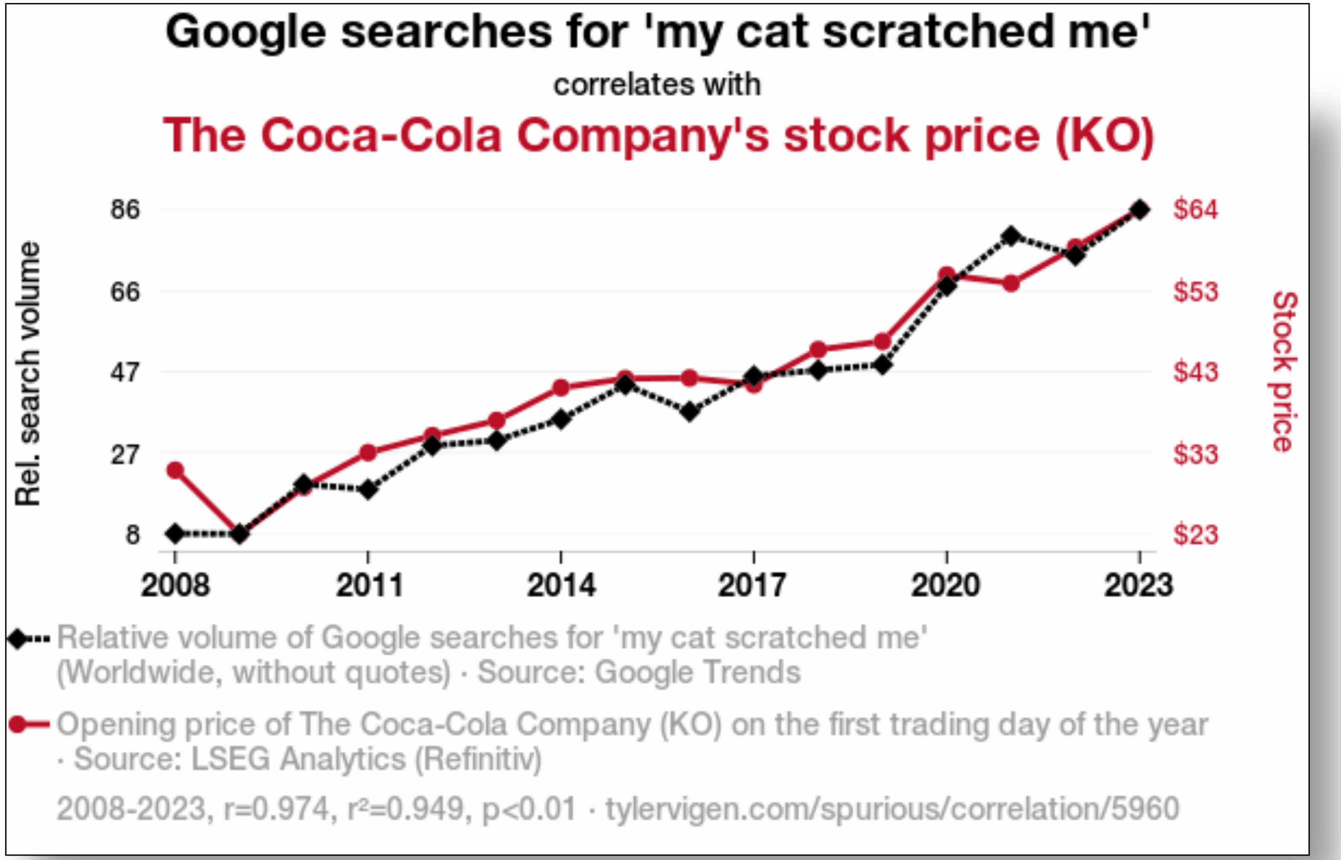
- Does changing variable x **cause** variable y to change (or vice versa)?

Limitations of Correlation



Source: Spurious Correlations - www.tylervigen.com

Limitations of Correlation



Note:

- Different y axes
- Y axes don't start at 0
- Different rate of change of y axes

Found via 'dredging' lots of variables, looking for a high correlation coefficients.

Source: Spurious Correlations - www.tylervigen.com

Limitations of Correlation

Naturally, there is probably little connection between cat scratches and the popularity of a soft drink.

BUT the saying *correlation does not imply causation* is not simply limited to these spurious correlations.

Let’s take another (this time, made up) example.

Limitations of Correlation

Let’s say we notice a strong positive correlation between **temperature data** and **ice cream sales** – i.e. as the temperature rises, more ice cream is sold.

‘But *correlation does not imply causation.’*

Does that mean that we don’t know whether *[warmer weather leads to more ice cream sales]* or whether *[ice cream sales leads to warmer weather]*?

Limitations of Correlation

Naturally, we know ice cream sales don't cause warmer weather, but *'correlation != causation'* still applies.

It *could* be the case that there is a *direct* connection between temperature and ice cream sales.

but it also *could* mean that there is an *indirect* connection between these variables.

Limitations of Correlation

Direct correlation is where two or more variables are directly connected:

- [Temperature] > [Ice cream sales]

Indirect correlation is where two variables are not directly connected, but there is one or more variable that is directly connecting them:

- [Temperature] > [Number of people going to the beach] > [Number of beach-side ice cream stalls open for business] > [Ice cream sales]

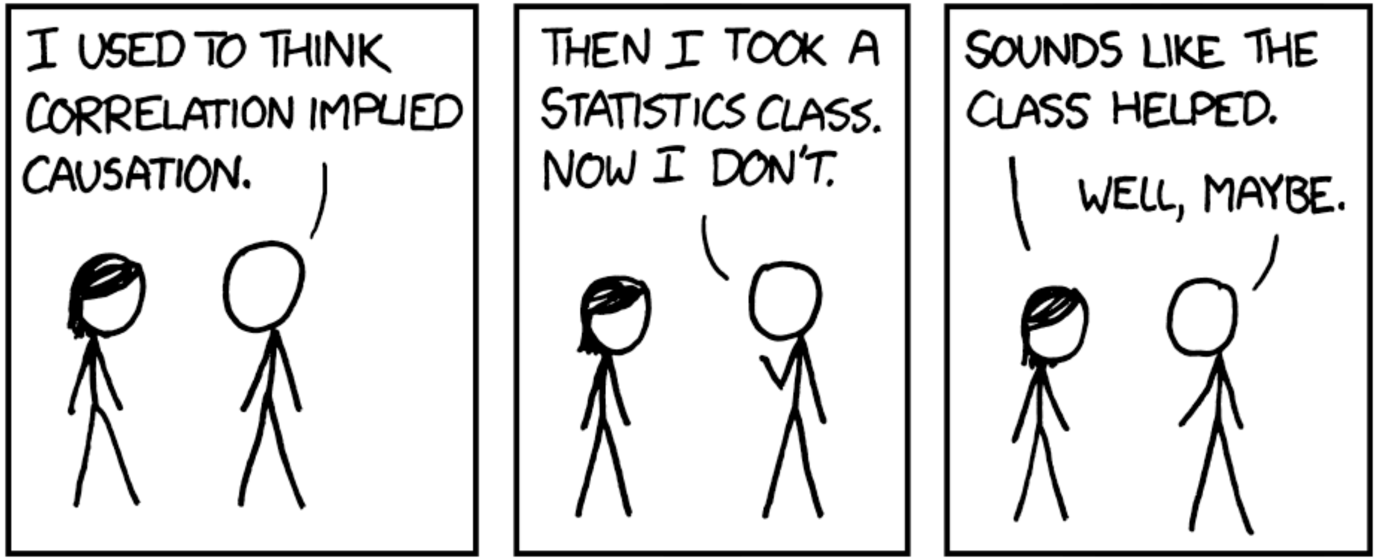
Limitations of Correlation

In other words, when variables **[A]** and **[B]** are correlated, it could be that:

- **[A]** and **[B]** are completely unconnected (spurious)
- **[A]** causes **[B]** (*directly*)
- **[B]** causes **[A]** (*directly*)
- **[A]** causes **[D]**, which causes **[C]**, which causes **[B]** (*indirectly*)
- **[E]** causes **[A]** and **[B]** (*confounding factor*)
- Etc…

Limitations of Correlation

The point is that we can’t say for sure. Correlation can only tell us so much.



Source: https://xkcd.com/552/

Limitations of Descriptive Statistics in General

Limitations of Descriptive Statistics

As a data scientist, descriptive statistics are invaluable.

But they are only a subset of the tools you have at your disposal!

Imagine We Have a Dataset With...

Imagine we have a dataset with:

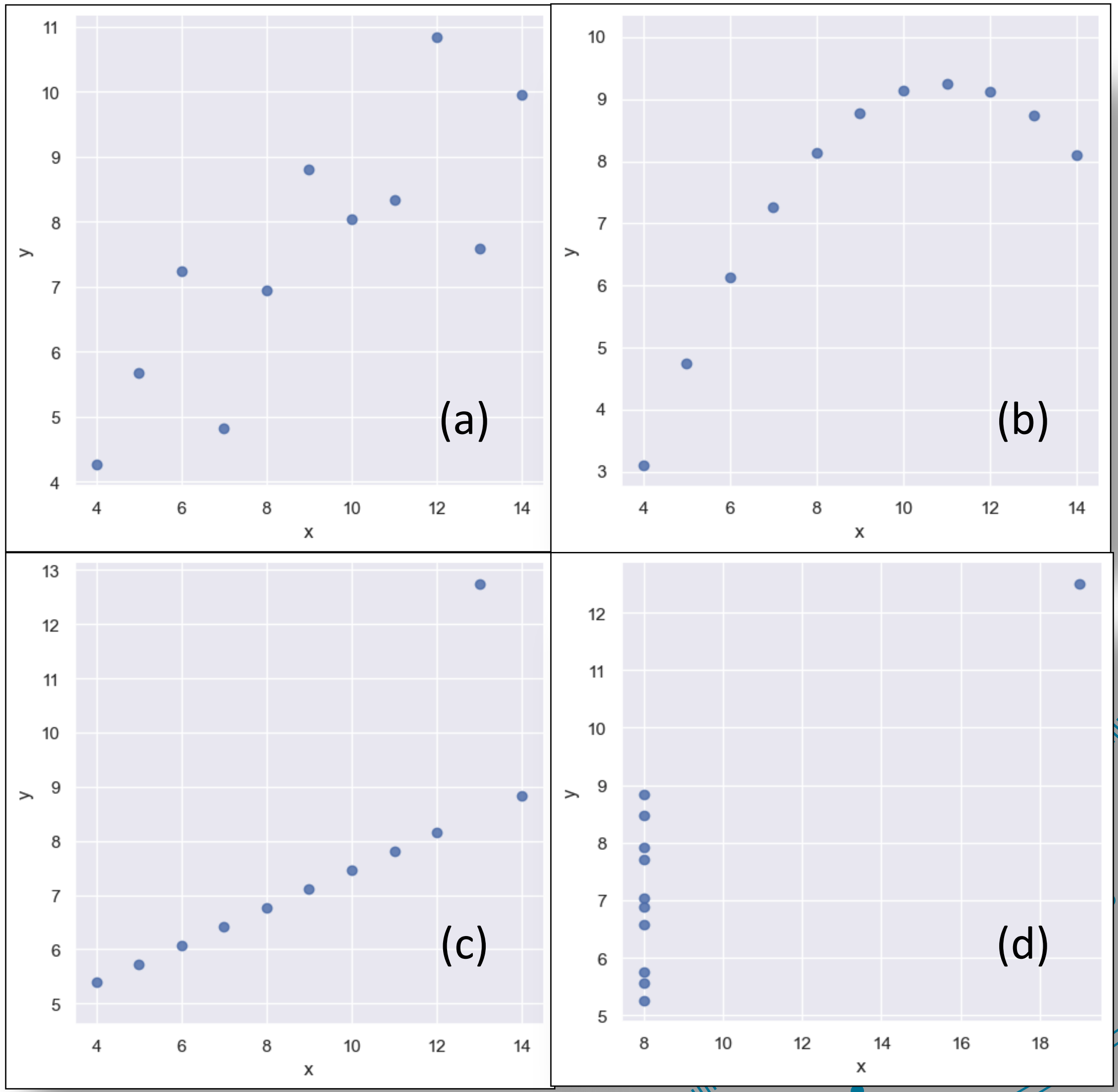
$$\bar{x} = 9$$

$$s_x = 3.32$$

$$\bar{y} = 7.5$$

$$s_y = 2.03$$

$$r_{xy} = 0.816$$



What does that look like?

Imagine We Have a Dataset With…

<u>Take a minute to guess which one it is:</u>

(a); (b); (c); or (d)

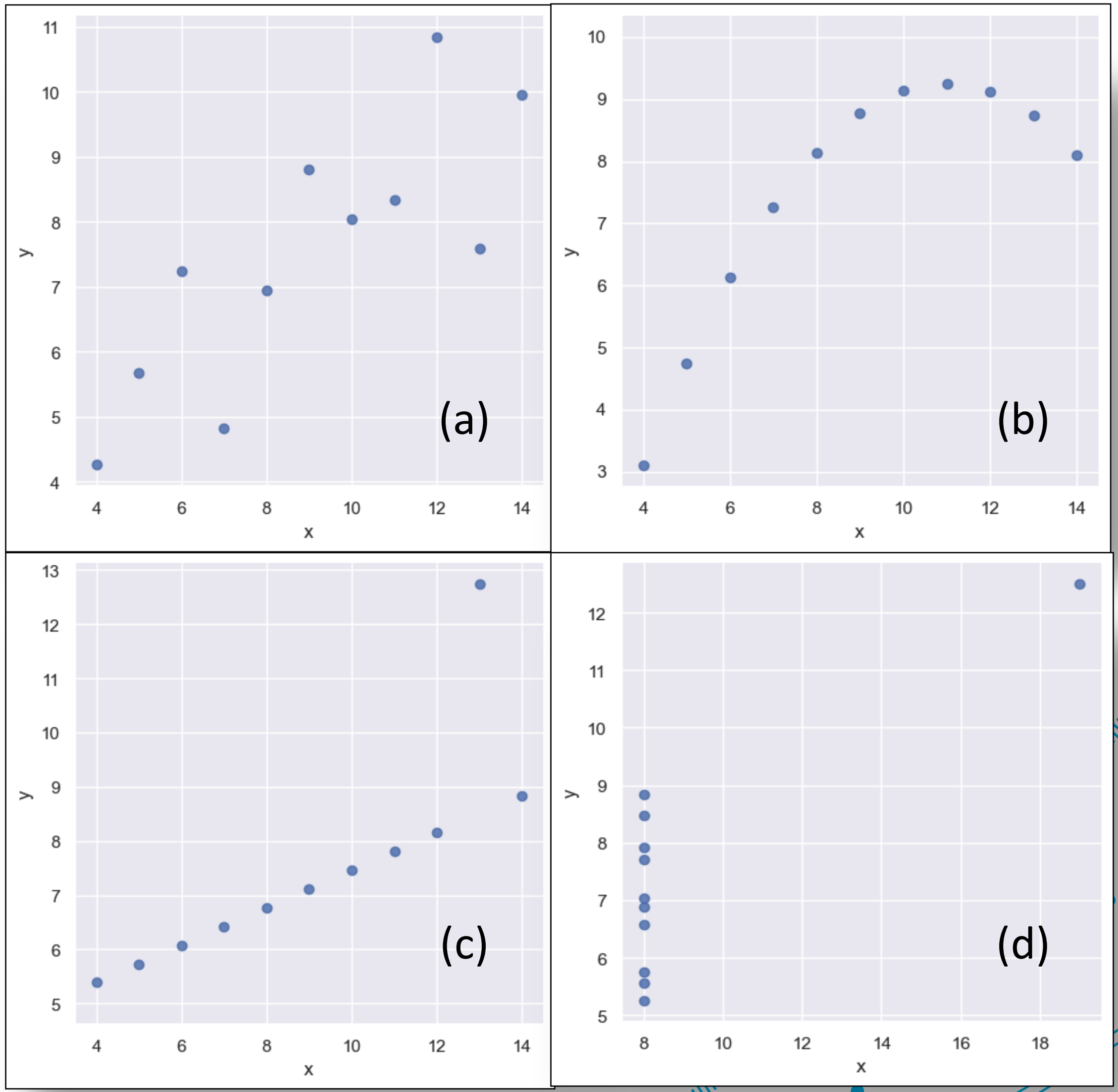
$$\bar{x} = 9$$

$$s_x = 3.32$$

$$\bar{y} = 7.5$$

$$s_y = 2.03$$

$$r_{xy} = 0.816$$



Anscombe's Quartet

Answer: **All of the above!**

All four of these datasets have:

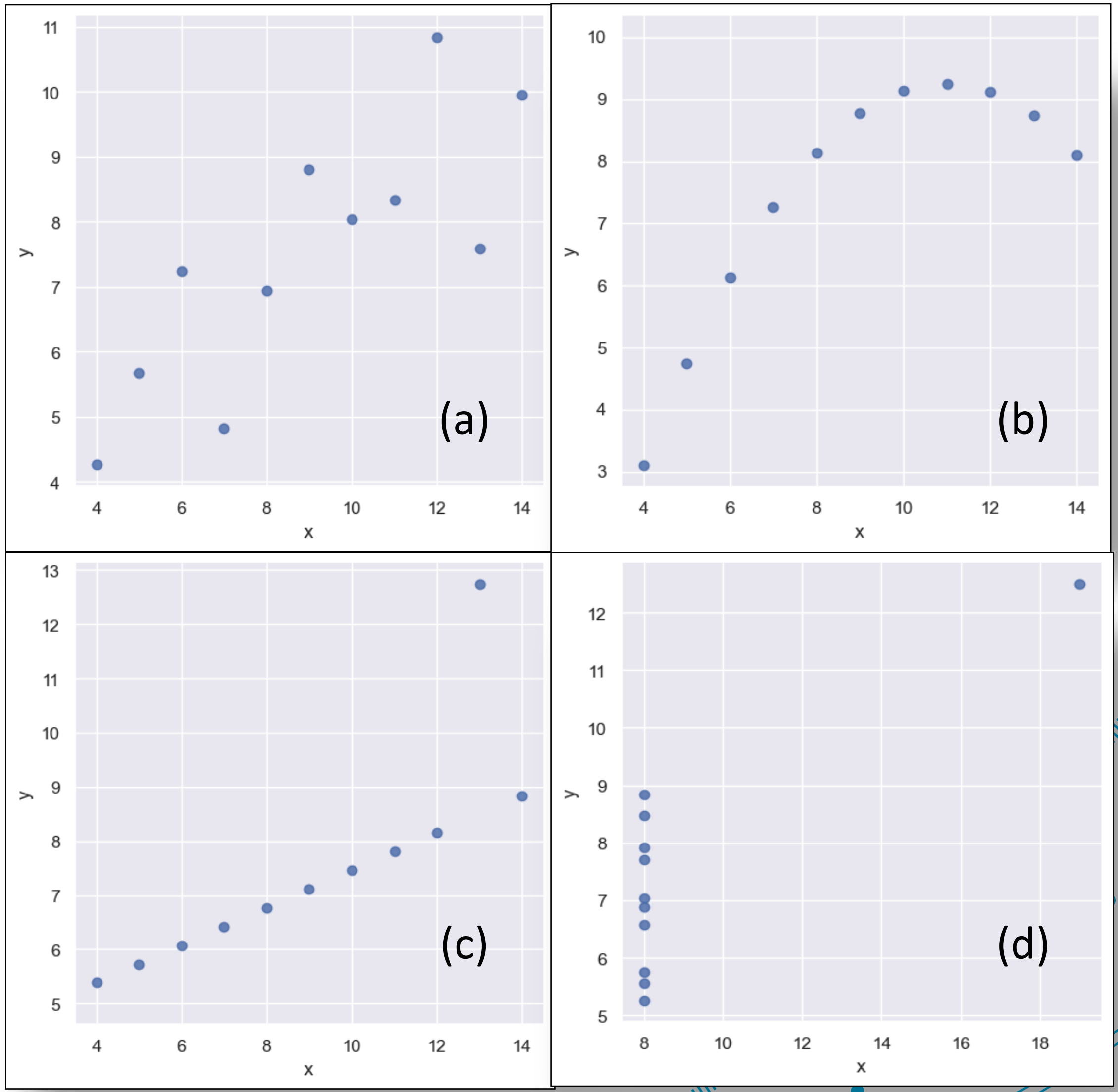
$$\bar{x} = 9$$

$$s_x = 3.32$$

$$\bar{y} = 7.5$$

$$s_y = 2.03$$

$$r_{xy} = 0.816$$



Descriptive Statistics vs Visualisations

Descriptive statistics give a 'partial picture' about your datasets.

Data visualisations can also play a vital role in helping you to gain more information and understanding about your data!