EXPLORE AI ACADEMY

Identifying patterns

Evaluating your line

How to evaluate the line of best fit

We can use various tools to **analyse** the line of best fit and **evaluate** its ability to represent the data. The **three tools** we will focus on are:

Correlation coefficient	Measures the strength and direction of a linear relationship between two variables.
R-squared (R^2)	Measures how well a line of best fit fits the data.
Heatmaps	A visualisation of the relationship between two or more variables.

Identifying patterns

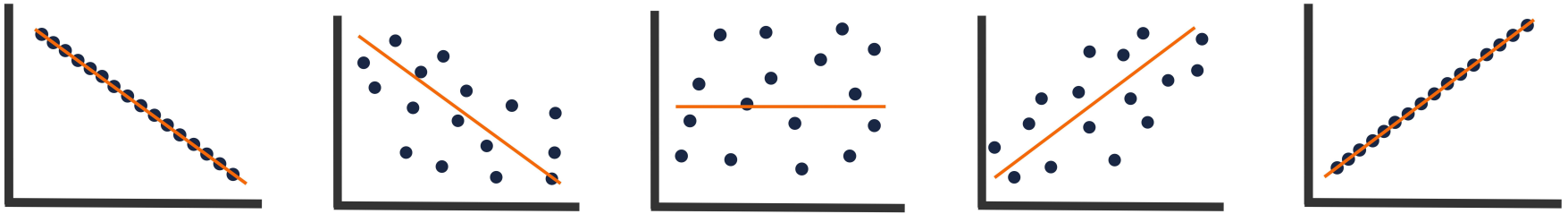
What is a correlation coefficient?

The correlation coefficient is a measure of the **strength** and **direction** of the **linear relationship between two variables**. It can be positive or negative and **ranges from -1 to +1.**

As the independent variable **increases** the dependent variable **decreases**.

As the independent variable **increases** the dependent variable **increases**.

-1 -0.5 0 +0.5 +1



Correlation function in Google Sheets

The **CORREL** function in Google Sheets is used to find the **correlation coefficient between two arrays.**

CORREL syntax:

```
=CORREL(data_y, data_x)
```

- **`data_y`** – The range representing the array or matrix of dependent data.
- **`data_x`** – The range representing the array or matrix of independent data.

Sample usage: `CORREL(A2:A100,B2:B100)`

What is R^2?

R^2 measures the **goodness of fit** of a line and can help to **evaluate how well the line fits the data**.

It tells us how much of the **variation in the dependent variable is explained by the independent variable**.

It **ranges from 0 to 1**, with 0 indicating the line does not fit the data at all and 1 indicating a perfect fit.

Interpreting R^2

> R^2 is calculated by **dividing the explained variance by the total variance in the dependent variable**, resulting in a value ranging from **0 to 1**.

Formula:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where **RSS** is the **sum of squares of residuals** and **TSS** is the **total sum of squares**.

$R^2=1$

A value of 1 indicates a perfect fit, i.e. all of the variation in the dependent variable is explained by the independent variable.

$0<R^2<1$

Values between 0 and 1 indicate the proportion of variation in the dependent variable that is explained by the independent variable.

$R^2=0$

A value of 0 indicates that the line of best fit does not fit the data at all.

R^2 function in Google Sheets

R^2 can be found in two ways in Google Sheets: using the **RSQ()** function and using the **chart editor**.

RSQ() syntax

```
=RSQ(data_y, data_x)
```

- **data_y** – The range representing the array or matrix of dependent data.
- **data_x** – The range representing the array or matrix of independent data.

Sample usage: `RSQ(A2:A100,B2:B100)`

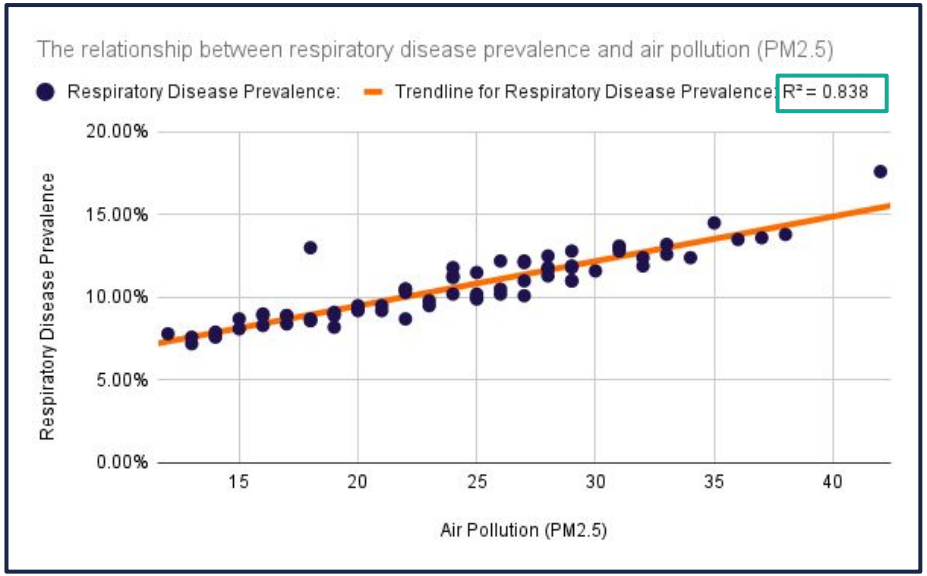
Chart editor

- In the chart editor, click on the **Customize** tab.
- Click the **Series** section.
- Scroll down and check the box next to **Show R^2**.
- The **R^2** value will appear on the chart.

Identifying patterns

Understanding a high R^2 value

Let's explore how R^2 can be applied to **assess the relationship between air pollution and respiratory disease prevalence** in African cities.



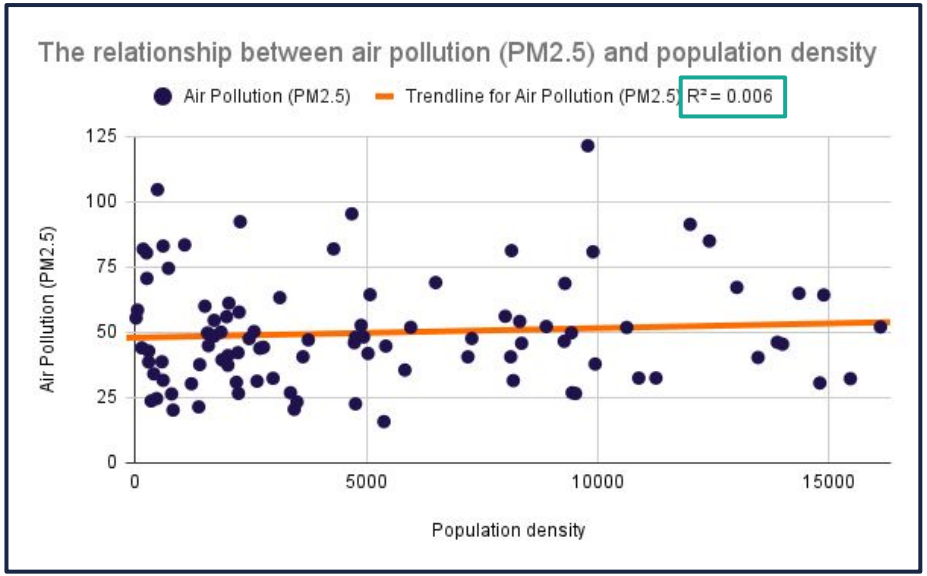
Analysis:

- We have a high R^2 value of **0.838**, indicating **the line is fitting the data well**.
- This means that **83.8% of the variability** in respiratory disease prevalence **can be explained** by air pollution levels.
- The remaining **16.2% could be attributed to other factors** not accounted for in our analysis.
- The high R^2 value suggests that air pollution **plays a significant role** in respiratory disease prevalence.

Identifying patterns

Understanding a low R^2 value

Let's explore how R^2 can be applied to **assess the relationship between population density and air pollution levels** in African cities.



Analysis:

- We have a low R^2 value of **0.006**, indicating **the line is not a good fit for the data**.
- This means that **only 0.6% of the variability** in air pollution **can be explained** by population density.
- **Almost all** of the variability (99.4%) is **attributed to other factors** not considered in our analysis.
- The low R^2 value suggests that population density may **not be a strong predictor** of air pollution levels.

Identifying patterns

Heatmaps

> Heatmaps are a **visual** way to **analyse the relationship** between multiple variables.

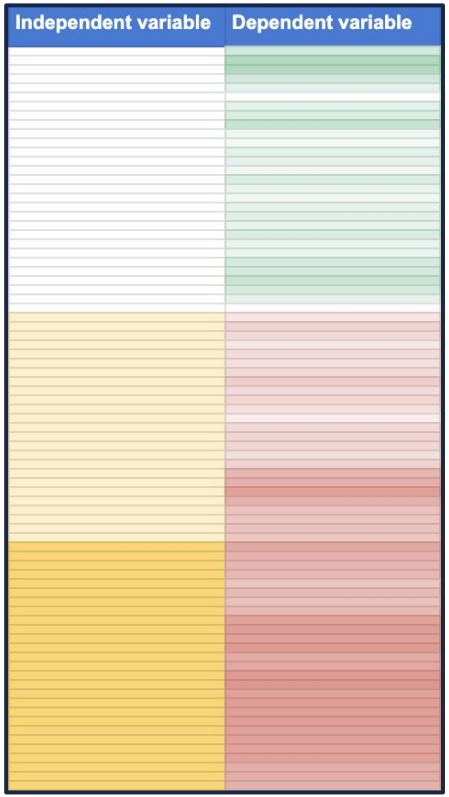
Heatmaps can help to **identify patterns** in the data that **may not be obvious** from a scatter plot.

They use **colour** to represent the value of a cell. In Google Sheets we do this with **conditional formatting**.

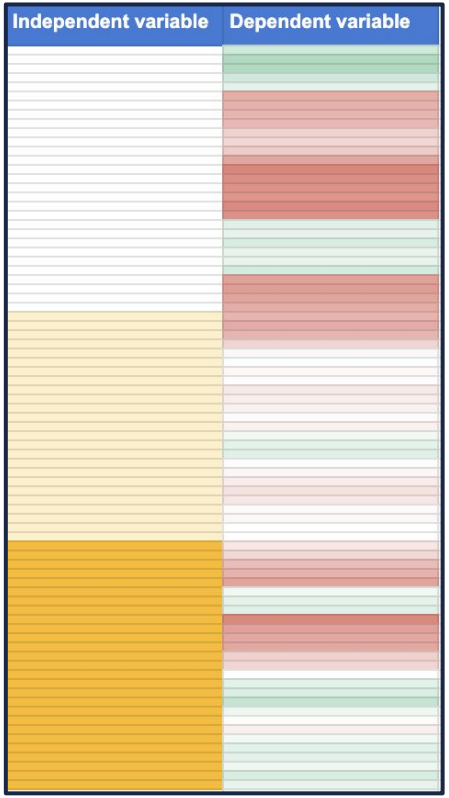
A	B	C
latitude	price	
0.516667	0.2651	
0.516667	0.261	
0.516667	0.2242	
0.516667	0.2885	
0.516667	0.3198	
0.516667	0.2952	
0.516667	0.3285	
1.750845	0.461	
1.750845	0.4627	
1.750845	0.4726	
1.750845	0.4684	
1.750845	0.4578	
1.750845	0.4587	
1.750845	0.4439	
2.33025	0.3747	
2.33025	0.3737	
3.11904	0.461	
3.11904	0.6478	
3.11904	0.4726	
3.11904	0.3945	

Interpreting the heatmap

If the heatmap shows a **gradual colour** change for the dependent variable as the independent variable changes, it is likely that there **is a linear relationship** between the two variables.



If the heatmap shows a **random distribution of colours** without any noticeable pattern or trend, there is **no linear relationship** between the two variables.



Tools to evaluate the line of best fit – summary

Correlation	R^2	Heatmaps
Measures the **strength and direction** of the linear relationship between two variables.	Measures the **proportion of the variance** in the dependent variable that is explained by the independent variable(s).	Provides a **visual representation of the relationship** between two variables.
Useful for **quantifying the strength** of the line of best fit.	Useful for **quantifying the goodness of fit** of the line of best fit.	Useful for **visually identifying patterns and trends** in the data.
If the relationship is **not linear** it **may not accurately reflect the relationship** between the variables.	If the relationship is **not linear**, it may **not accurately reflect the quality** of the line of best fit.	If the relationship is **not linear**, it **can still be used to detect other patterns**.