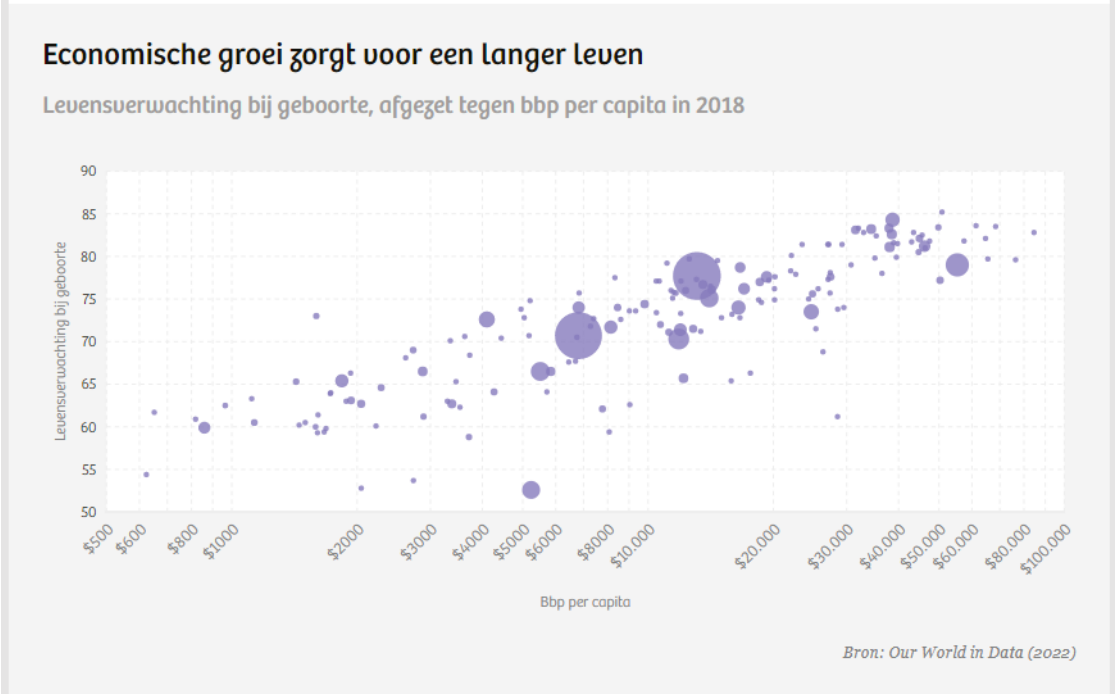
# Correlatie en lineaire regressie

www.karelappeltans.be

April 16, 2024

## 1 Inleiding




## 2 puntenwolk

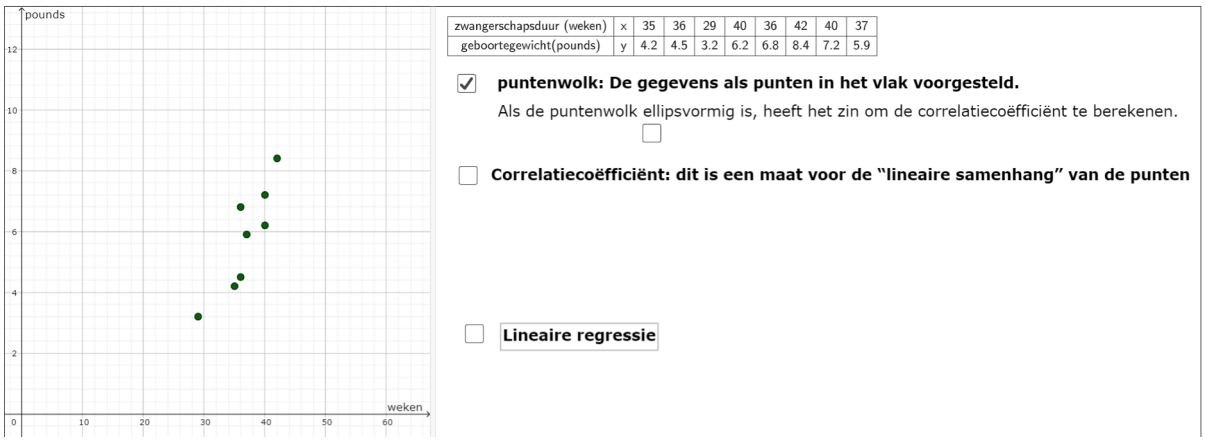



Figure 1: https://www.geogebra.org/m/ywda8eSq

# 3 correlatiecoëfficiënt (van Pearson)

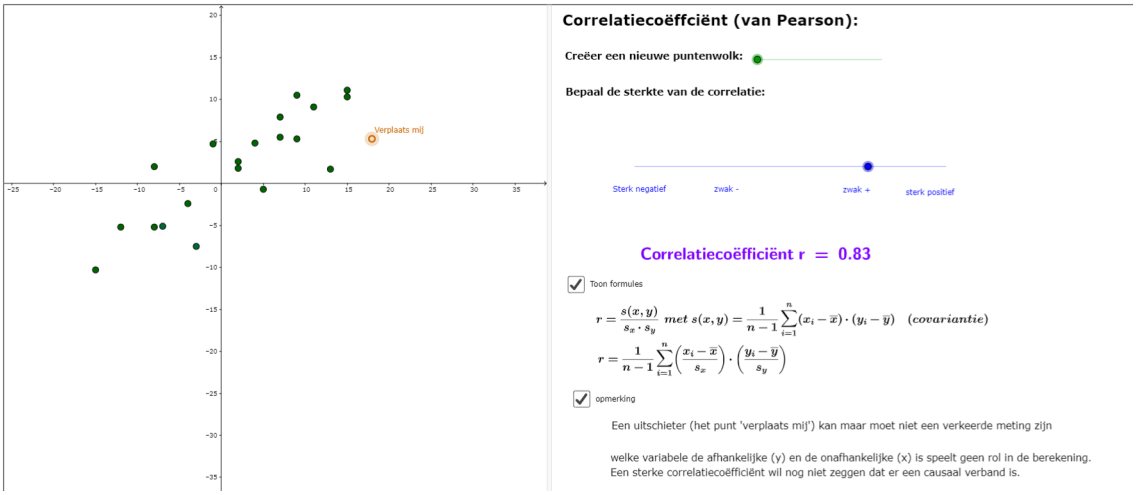



Figure 2: https://www.geogebra.org/m/ywda8eSq

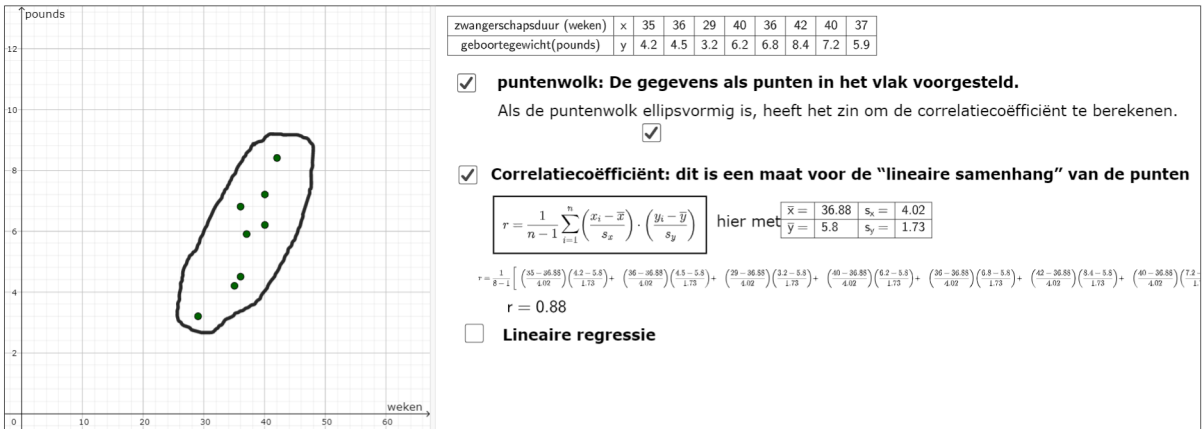
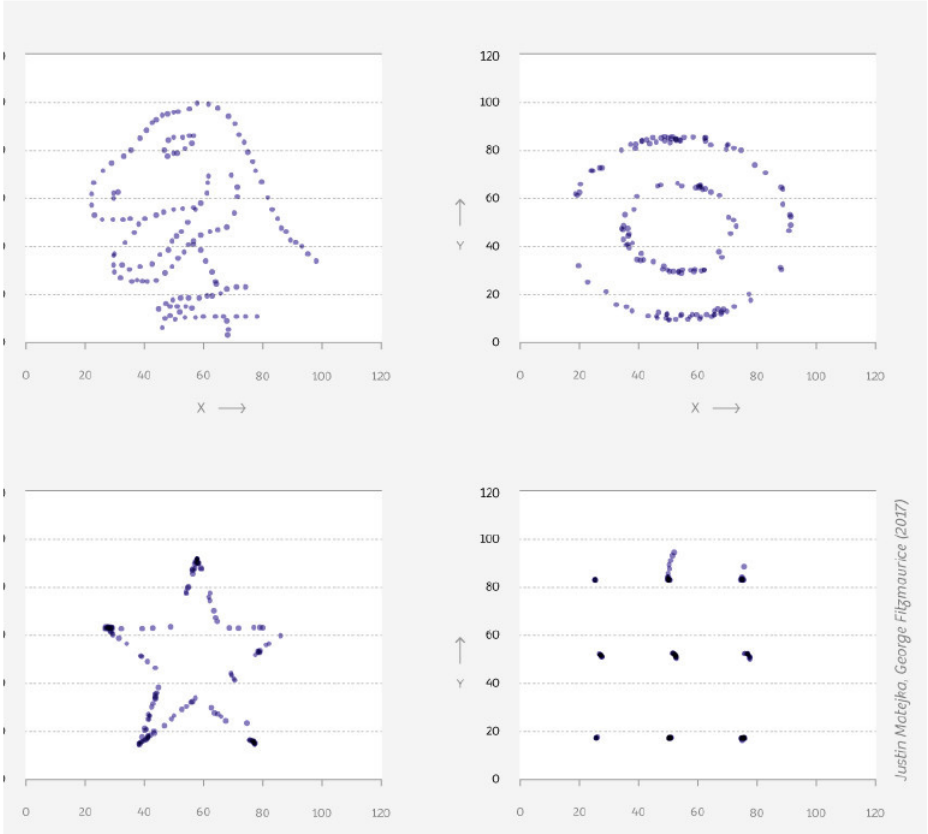



Figure 3: https://www.geogebra.org/m/ywda8eSq

Dat het belangrijk is om eerst de puntenwolk te tekenen mag blijken uit onderstaande voorbeelden:




Al deze 'puntenwolken' hebben voor de x- en y-variabele hetzelfde gemiddelde en standaardafwijking en hebben dezelfde correlatiecoëfficiënt.

# 4 lineaire regressie

## 4.1 vergelijking rechte

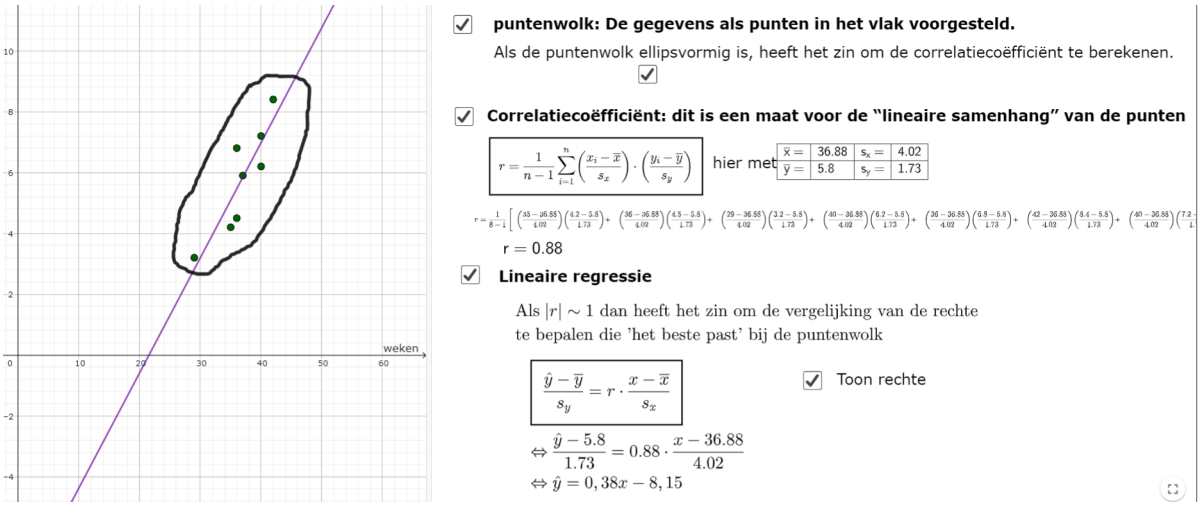



Figure 4: https://www.geogebra.org/m/ywda8eSq

## 4.2 interpolatie en extrapolatie

Interpolatie is het proces waarbij de best passende lijn wordt gebruikt om de waarde van de ene variabele te schatten op basis van de waarde van een andere, op voorwaarde dat de waarde die u

gebruikt binnen het bereik van uw gegevens ligt. Als het buiten het bereik valt, zou je Extrapolatie gebruiken. Bij extrapolatie let op of de berekeningen nog wel zin hebben. Bijvoorbeeld, bereken het verwacht geboortegewicht bij een zwangerschapsduur van 39 weken (interpolatie) respectievelijk 45 weken (extrapolatie).
Oplossing $\hat{y} = 0,38 \cdot 39 - 8,15 = 6,67$. Voor 45 weken heeft dit geen zin.

# 5 kleinste kwadratenmethode

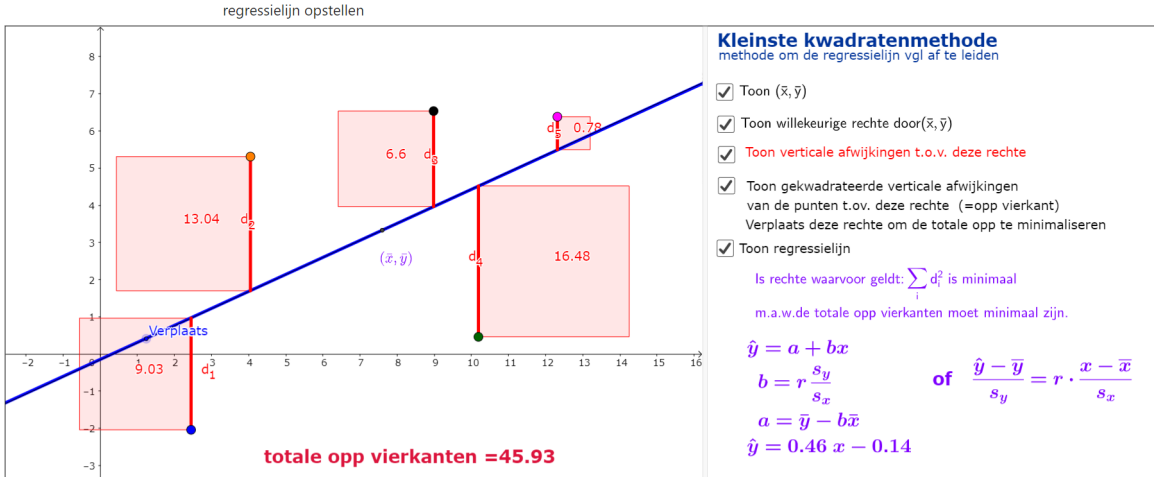
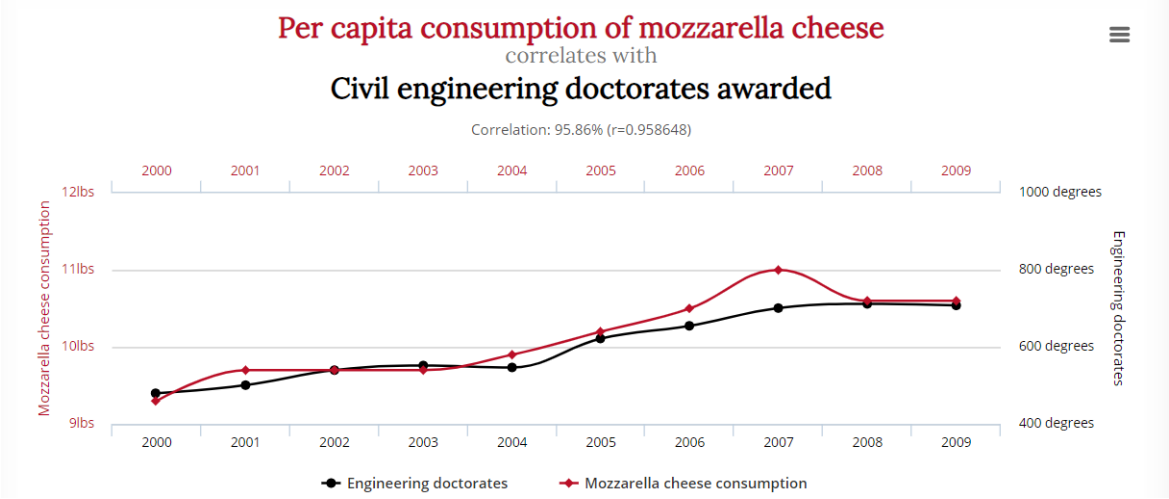
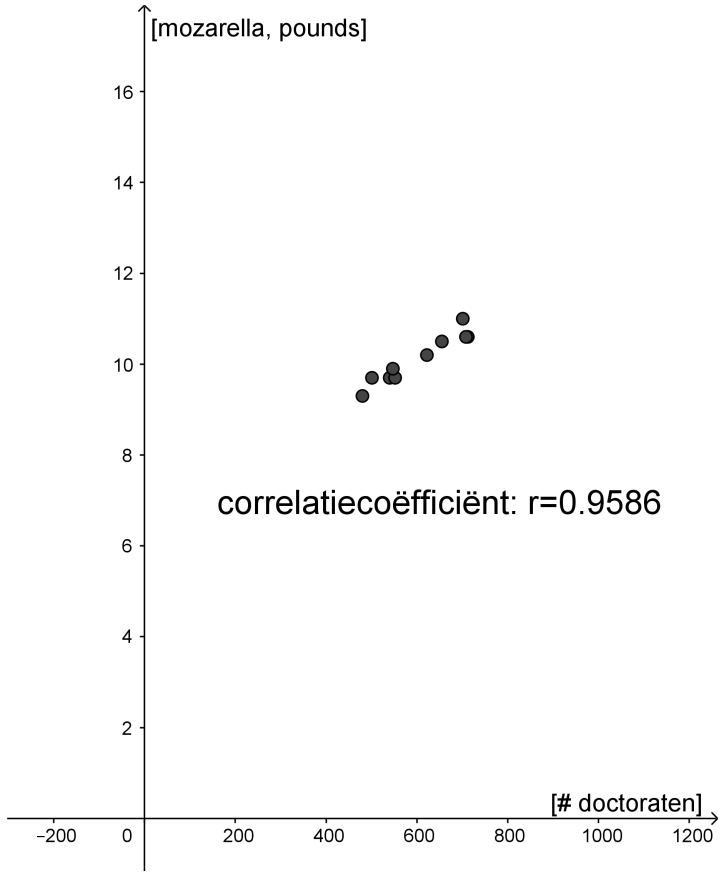



Figure 5: https://www.geogebra.org/m/ywda8eSq

# 6 correlatie is geen causaliteit






[http://www.tylervigen.com/spurious-correlations]http://www.tylervigen.com/spurious-correlations

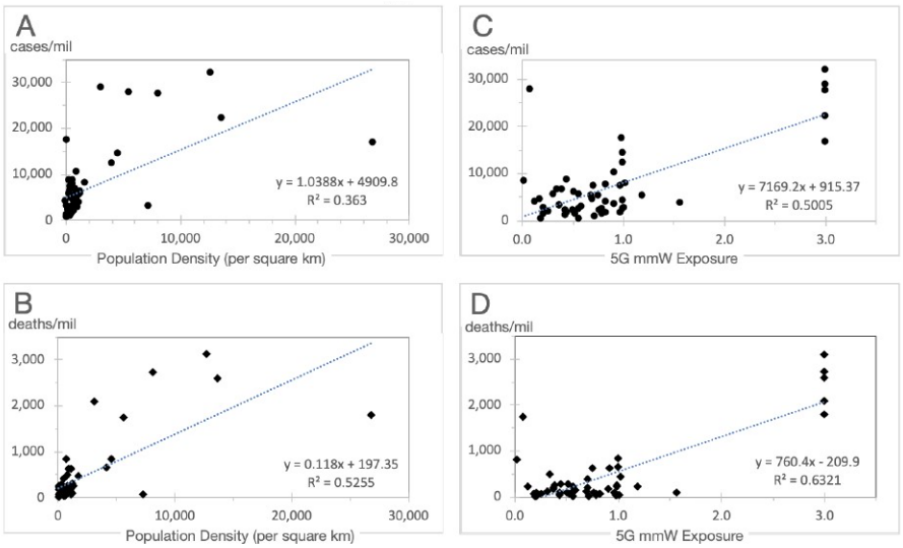
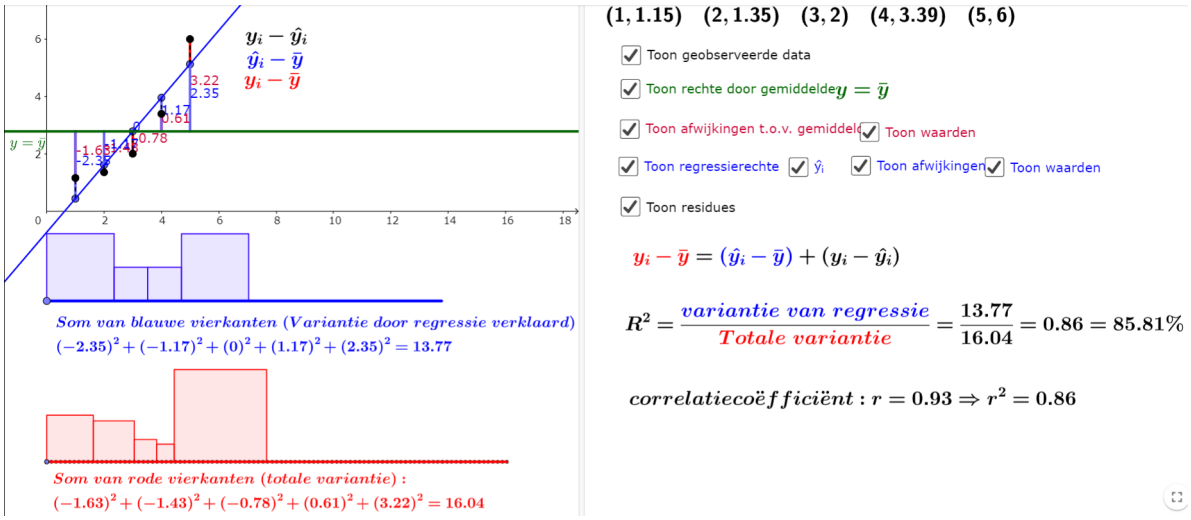



***Figure 6.*** *Regression plot for 53 counties with 5G mmW technology for COVID-19 attributed cases/million and deaths/million as a function of population density (A & B) and as a function of 5G mmW exposure (C & D) through May 31, 2020.*

# 7 determinatiecoëfficiënt




[https://www.geogebra.org/m/ywda8eSq]https://www.geogebra.org/m/ywda8eSq

# 8 andere notaties

In andere handboeken kan je nog andere formules en notaties aantreffen. Deze worden hier niet besproken, behalve de covariantie

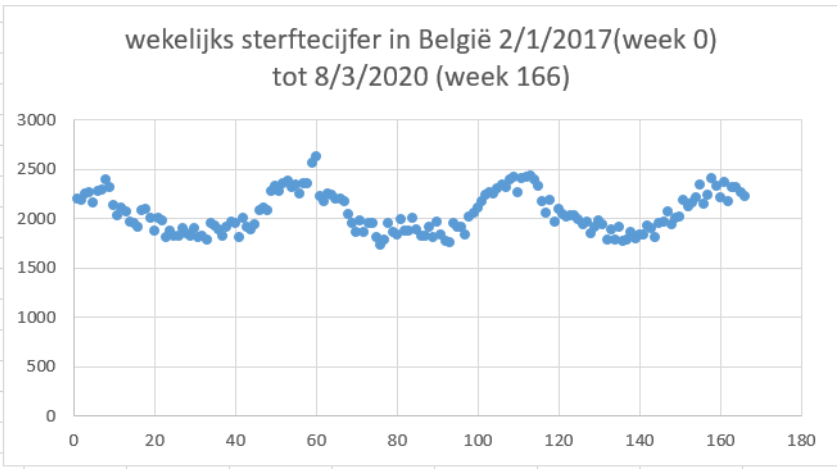
$$covariantie: \ s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

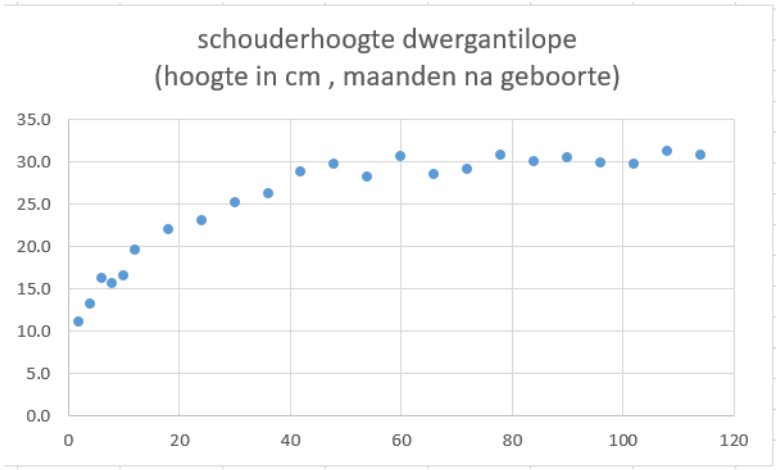
Dit geeft dan voor de correlatiecoëfficiënt:

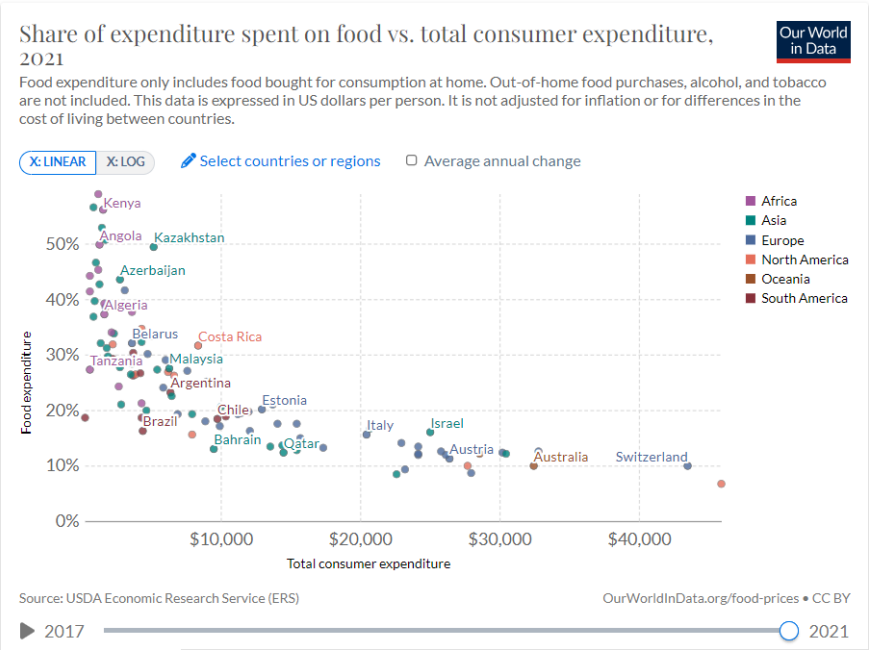
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

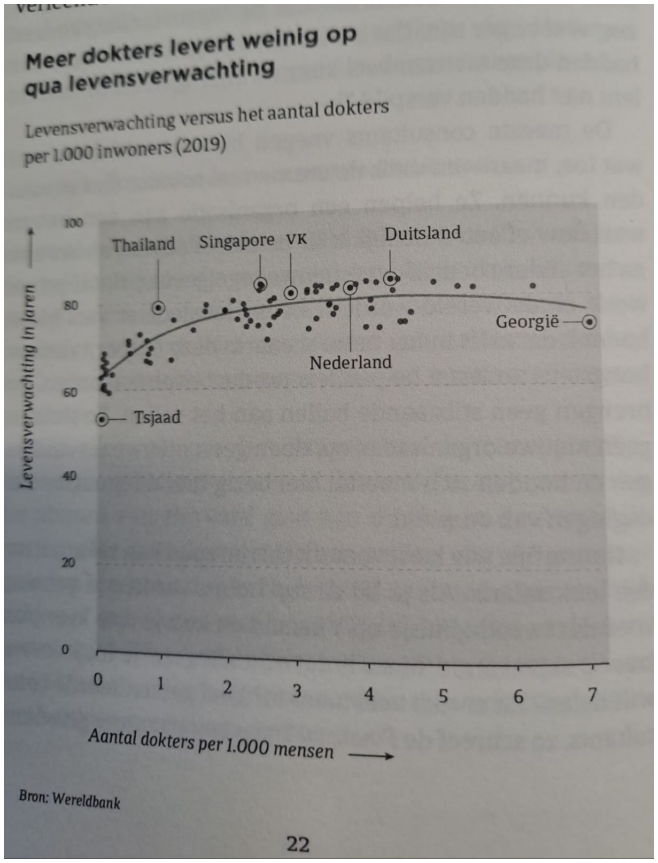
# 9 andere regressievormen

Niet alle puntenwolken vertonen een lineair verband. Enkele voorbeelden. Welk verband herken je?




schouderhoogte dwergantilope
(hoogte in cm , maanden na geboorte)
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
20
40
60
80
100
120


Share of expenditure spent on food vs. total consumer expenditure, 2021
Our World in Data
Food expenditure only includes food bought for consumption at home. Out-of-home food purchases, alcohol, and tobacco are not included. This data is expressed in US dollars per person. It is not adjusted for inflation or for differences in the cost of living between countries.
X: LINEAR
X: LOG
Select countries or regions
Average annual change
Africa
Asia
Europe
North America
Oceania
South America
Kenya
Angola
Kazakhstan
Azerbaijan
Algeria
Belarus
Costa Rica
Tanzania
Malaysia
Argentina
Estonia
Chile
Brazil
Bahrain
Qatar
Italy
Israel
Austria
Australia
Switzerland
50%
40%
30%
20%
10%
0%
Food expenditure
$10,000
$20,000
$30,000
$40,000
Total consumer expenditure
Source: USDA Economic Research Service (ERS)
OurWorldInData.org/food-prices • CC BY
2017
2021




# 10 oefeningen

1. Een chemisch ingenieur onderzoekt het verband tussen de temperatuur en de opbrengst in een chemisch proces. Hij bekomt volgende gegevens

| T (°C) | 100 | 110 | 120 | 130 | 140 |
|---|---|---|---|---|---|
| opbrengst (%) | 45 | 51 | 54 | 61 | 66 |

   (a) Teken de puntenwolk

   (b) Bereken de correlatiecoëfficiënt

   (c) Bepaal de vergelijking van de regressierechte

   (d) Bereken de verwachte opbrengst bij een temperatuur van 122°C

2. Is er een lineair verband tussen de jaren rijervaring van een auto chauffeur en de maandelijkse kostprijs van de autoverzekering? Een onderzoeker bekomt de volgende gegevens van een random sample van 100 chauffeurs bij een bepaalde verzekeringsmaatschappij:

| | gemiddelde | standaardafwijking |
|---|---|---|
| jaren ervaring | 11,25 | 7,4 |
| maandelijkse premie | 69 | 11,8 |

   Er is ook gegeven dat $s_{xy} = -774,6$

   (a) Schets een mogelijke puntenwolk. Benoem zeker de assen!

   (b) Bereken de correlatiecoëfficiënt

(c) Bepaal de vergelijking van de regressierechte

(d) Bereken de verwachte premie voor een chauffeur met 10 jaar rijervaring

3. We willen een goedkoop appartement kopen in Hasselt (prijs van hoogstens 220000 euro) en vragen ons af of er een verband is tusen de prijs en de oppervlakte van dit type appartementen. We hebben daarom een aselect staal genomen van 50 'goedkope' Hasseltse appartementen en we hebben een aantal steekproef resultaten berekend:

| | gemiddelde | standaardafwijking |
|---|---|---|
| opp in $m^2$ | 71,92 | 22,33 |
| prijs in euro | 169242,22 | 38585,55 |

$s_{xy} = 644650,38$

(a) Bereken de sterkte van het lineaire verband tussen de oppervlakte en de prijs van deze 50 Hasseltse appartementen. Wat is je besluit? ($r = 0.7482$; vrij sterk positief verband)

(b) Schrijf zowel het theoretisch model als het geschatte model op als we de prijs willen voorspellen op basis van de oppervlakte. Interpreteer-als het zinvol is- de coëfficiënten van het geschatte model. (Prijs= a+b oppervlakte; $\widehat{prijs} = 76260.9551 + 1292.846opp$; het interpreteren van de constante in het geschatte model heeft geen zin: er zijn geen appartementen met oppervlakte (in de buurt van) 0 $m^2$. Het gaat hier om een extrapolatie van de gegevens; Als de oppervlakte met 1 $m^2$ toeneemt, dan zal de voorspelde gemiddelde prijs toenemen met 1292.85 euro)

(c) Bereken en interpreteer de determinatiecoëfficiënt van het geschatte model ($R^2 = 0.5598$. Dit betekent dat 56% van de spreiding in de prijzen kan verklaard worden door rekening te houden met de oppervlaktes van deze appartementen)

4. Slechts één van de volgende uitspraken is juist. Welke en verklaar waarom die uitspraak juist is en de andere uitspraken fout zijn.

(a) De correlatiecoëfficiënt tussen een voetballer zijn gewicht en de positie waar hij speelt is 0.54.

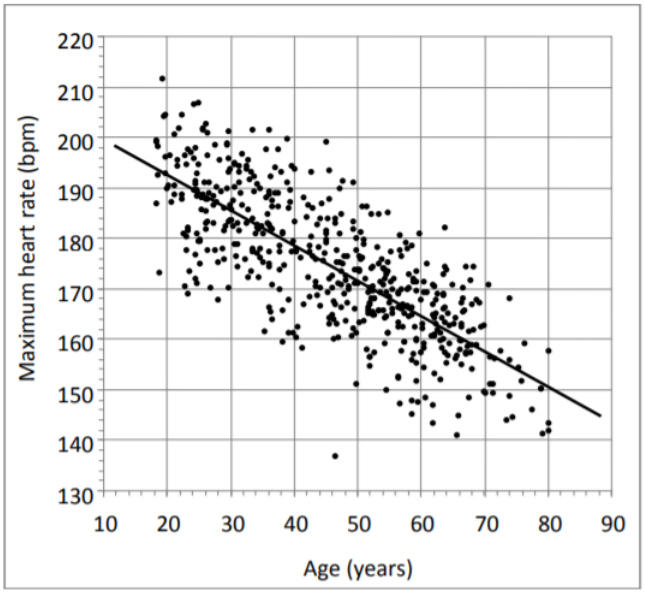
(b) De correlatiecoëfficiënt tussen de lengte van de auto en zijn verbruik is 0.71 kilometer per liter

(c) De correlatiecoëfficiënt tussen de verloning van een leerkracht wiskunde en de resultaten van zijn leerlingen bedraagt 0.42

(d) Er is een hoge correlatiecoëfficiënt van 1.09 tussen de lengte van een persoon en de spanwijdte van de armen

5. Bespreek onderstaande puntenwolk

A person's *maximum heart rate* is the highest rate at which their heart beats during certain extreme kinds of exercise. It is measured in beats per minute (bpm). It can be measured under controlled conditions. As part of a study in 2001, researchers measured the maximum heart rate of 514 adults and compared it to each person's age. The results were like those shown in the scatter plot below.




*Source:* Simulated data based on:Tanaka H, Monaghan KD, and Seals DR. *Age-predicted maximal heart rate revisited*, J. Am. Coll. Cardiol. 2001;37;153-156.

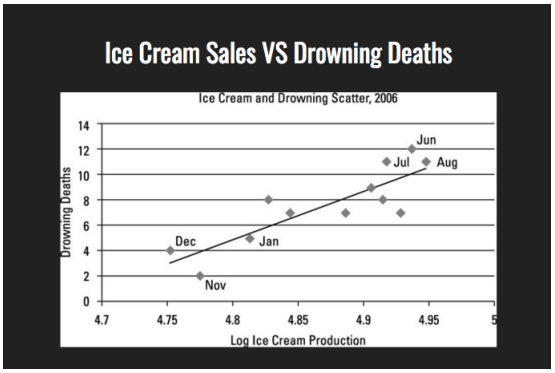
6. Verklaar volgende deelzin uit onderstaand artikel (gelezen op www.decorrespondent.nl): "maar hoe de causaliteit precies werkt, is minder makkelijk te zeggen"

Het onderzoek naar de psychologische effecten van sociale media op jongeren is even onvolgroeid als het object ervan

Nu is het wetenschappelijk onderzoek naar de psychologische effecten van sociale media op jongeren al even onvolgroeid als het object ervan. Het staat in de kinderschoenen en zowel Twenges data als haar cultuurpessimistische conclusies zijn onderwerp van kritiek geweest. Dat er namelijk een correlatie bestaat tussen smartphonegebruik en hoge depressiecijfers is wel duidelijk, maar hoe de causaliteit precies werkt, is minder makkelijk te zeggen.

*Meer feiten en minder hype, oordeelde The Guardian over Twenges boek*

7. Wat loopt hier mis?



8. Wat loopt hier mis?


**Figure 1** Prevalence of obesity and overweight among US adults: Observed during 1976–2004 and projected. The projected prevalence presented here are those based on our linear regression models.

Figure 6: https://onlinelibrary.wiley.com/doi/epdf/10.1038/oby.2008.351

9. Bespreek onderstaand spreidingsdiagram. Leg ook de betekenis van $R^2 = 0.5411$ uit.

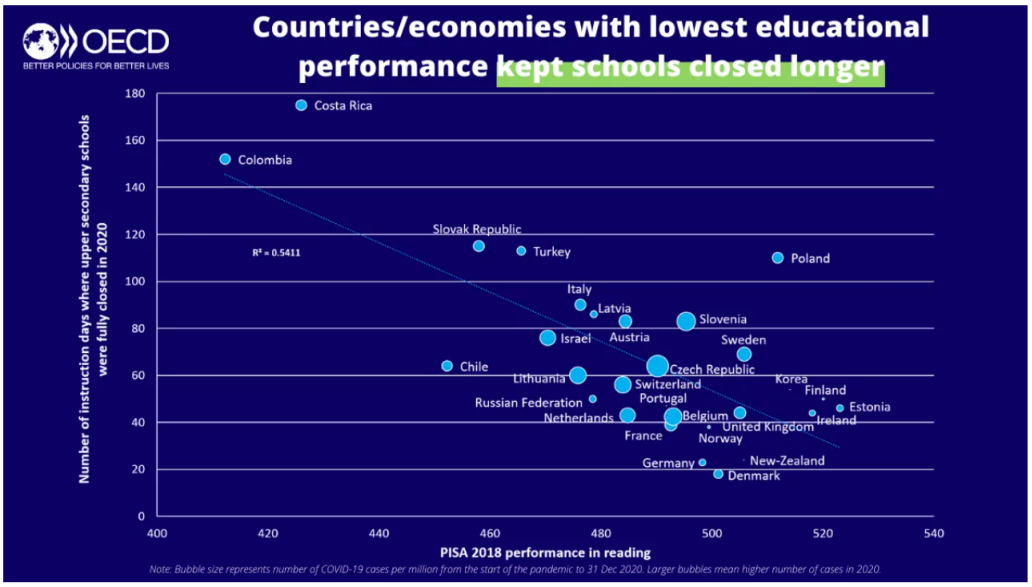
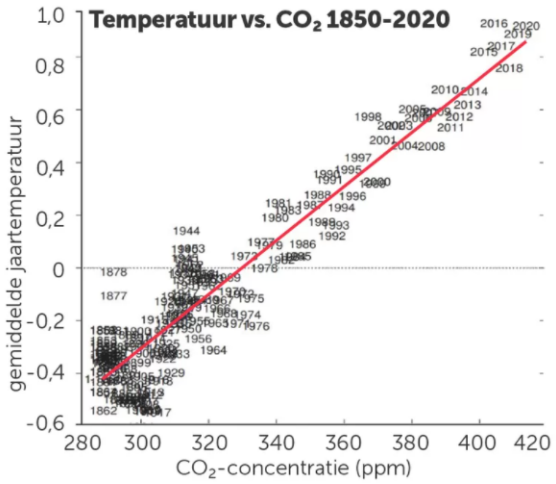



Figure 7: https://oecdedutoday.com/state-of-education-one-year-into-covid/

10. Bespreek onderstaand spreidingsdiagram, klimaattop Glasgow oktober 2021

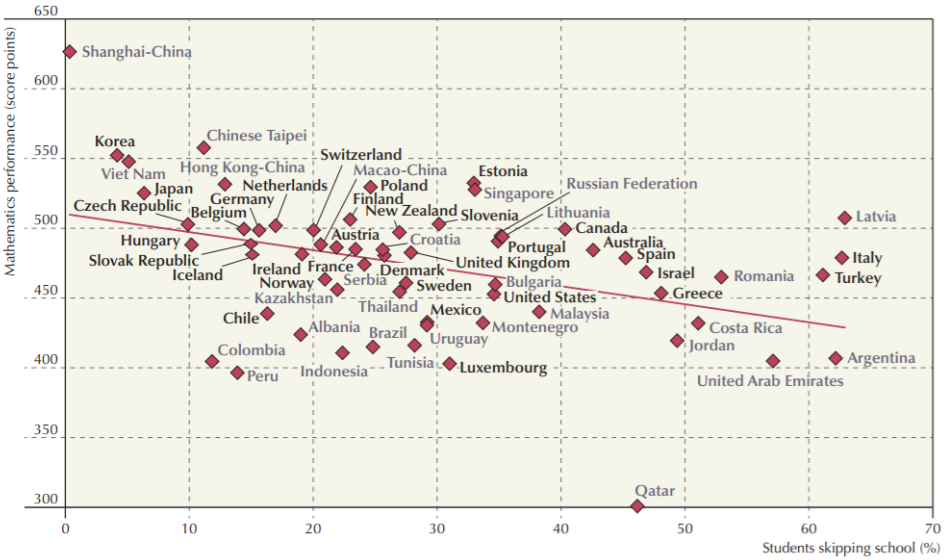
**4| HOE CO2 DE BOEL OPSTOOKT**




Beeld rv

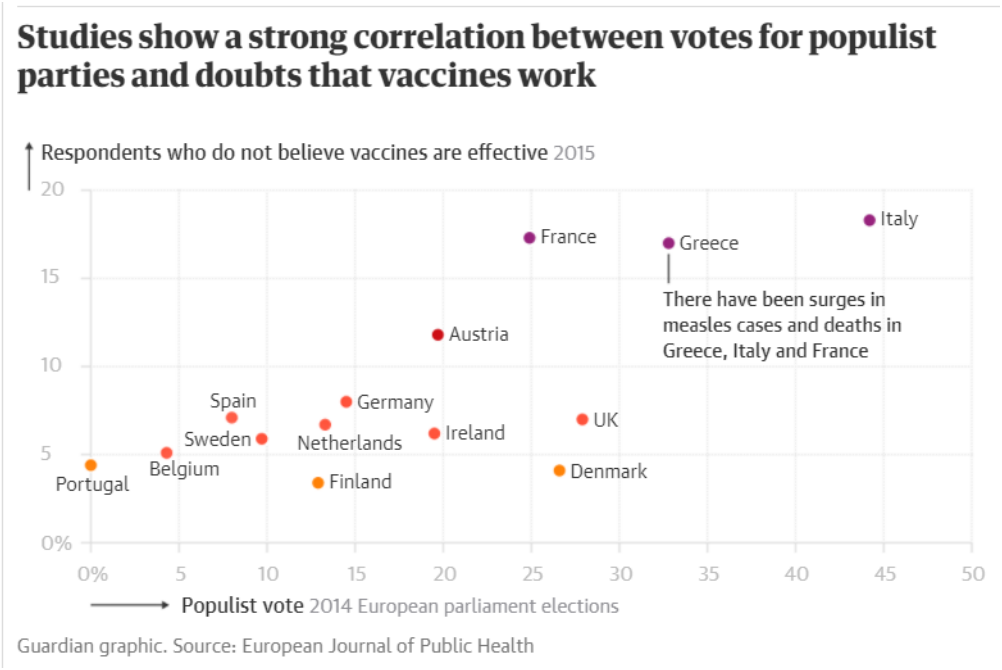
Talloos zijn de manieren om te zien dat CO2 achter de opwarming van de aarde zit. Voor de twijfelaars maakte KNMI-wetenschapper Geert Jan van Oldenborgh deze grafiek. Horizontaal zien we de hoeveelheid CO2 in de dampkring, verticaal de temperatuur van het jaar in kwestie. Voor fijnproevers: beide correleren met een 'r-waarde' van 0,939, een

11. Bespreek onderstaand spreidingsdiagram: (OECD,2014,p4)

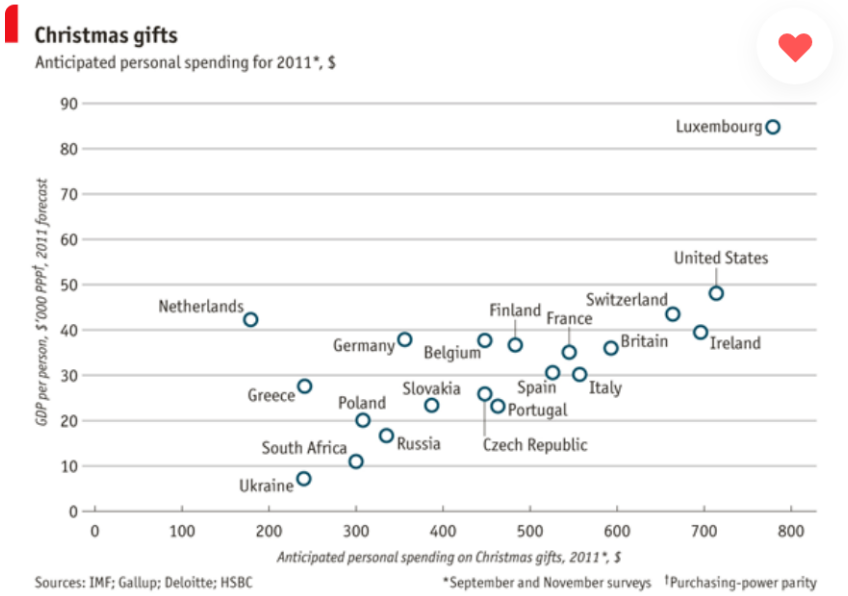



**Note:** Students skipping school refers to the percentage of students who had skipped a class or a day of school at least once in the two weeks prior to the PISA test.

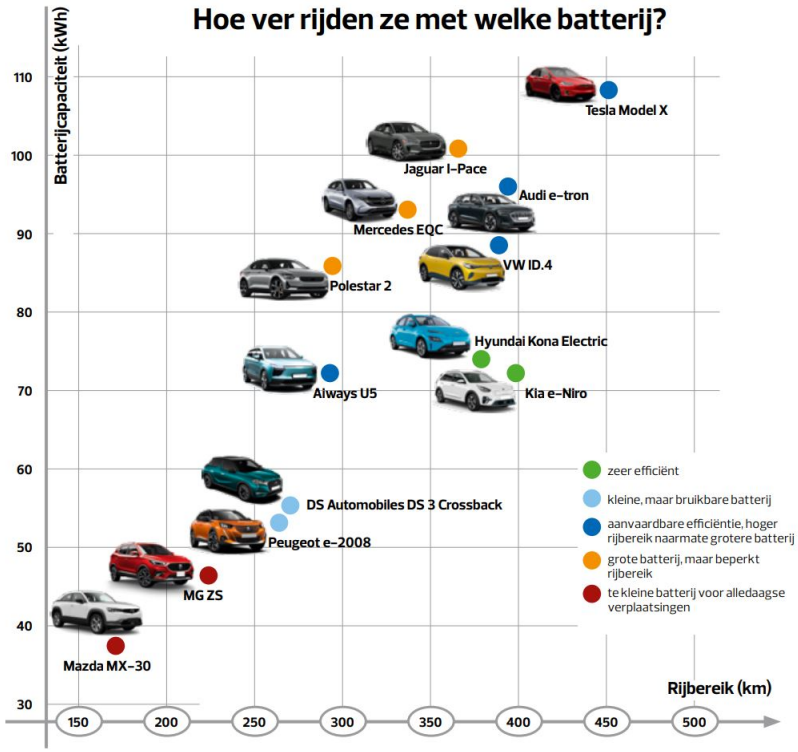
12. Bespreek onderstaand spreidingsdiagram:




13. Bespreek volgende puntenwolk (afkomstig uit The Economist, 12 december 2011):

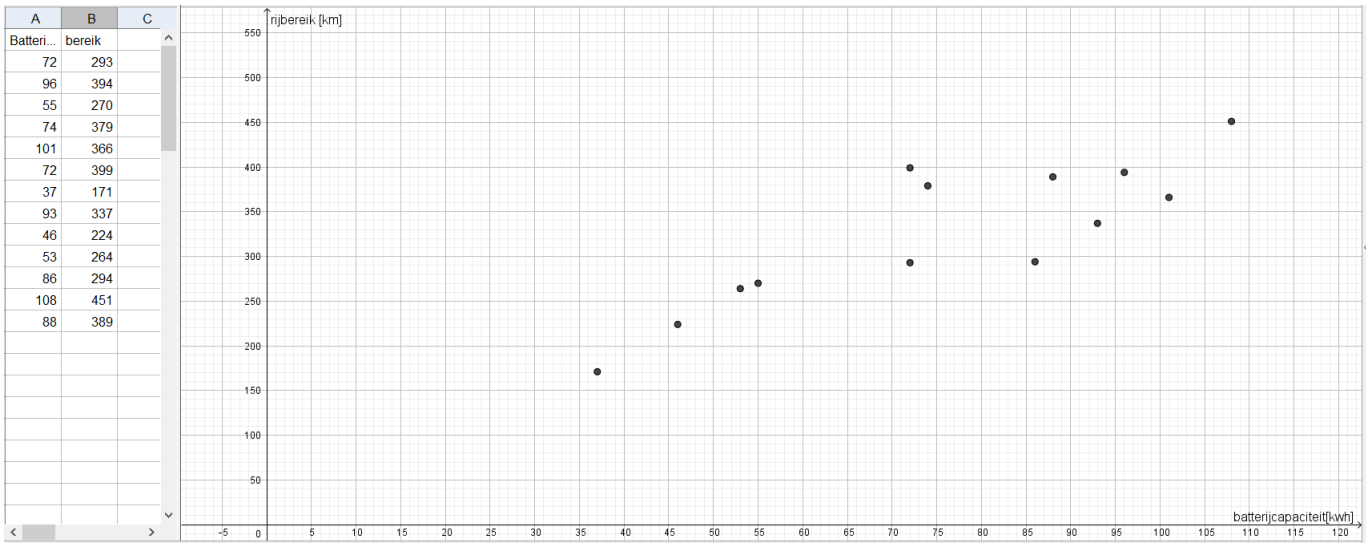



14. Gegeven onderstaande grafiek uit het tijdschrift testaankoop, september 2021




De batterijcapaciteit van elk model geeft weer hoeveel kWh er nodig is voor een volle batterij, dus verliezen tijdens het laden inbegrepen (zie tekst).

Bespreek deze grafiek. Wat zouden wij wiskundig anders doen?

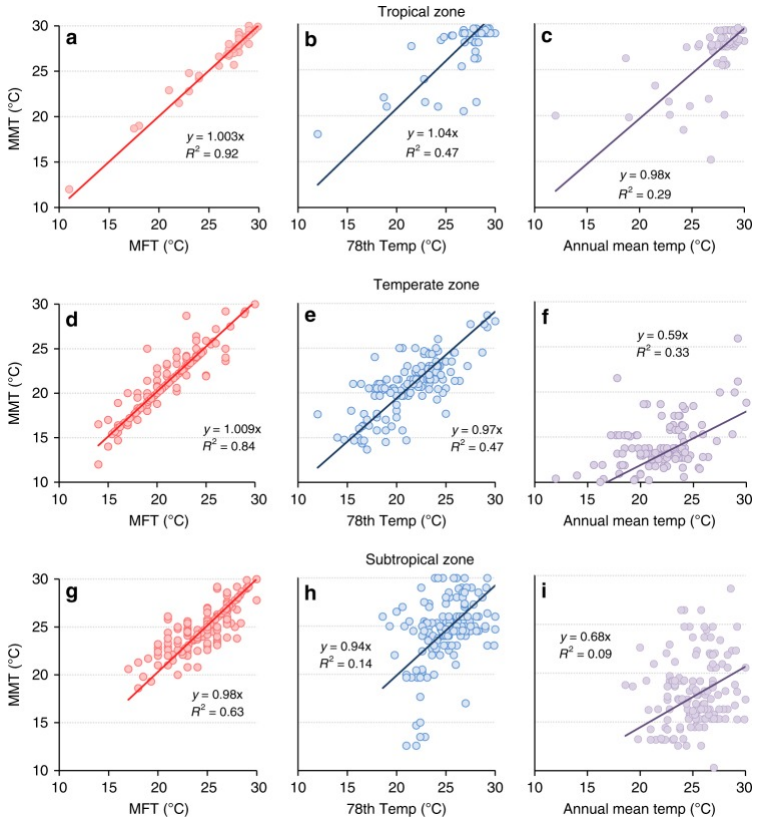


| A | B | C |
|---|---|---|
| Batteri... | bereik | |
| 72 | 293 | |
| 96 | 394 | |
| 55 | 270 | |
| 74 | 379 | |
| 101 | 366 | |
| 72 | 399 | |
| 37 | 171 | |
| 93 | 337 | |
| 46 | 224 | |
| 53 | 264 | |
| 86 | 294 | |
| 108 | 451 | |
| 88 | 389 | |



Bereken de correlatiecoëfficiënt en de vergelijking van de regressierechte
(A. $r = 0.85$ en $y = 3.07x + 93.99$)

15. De MMT (de minimum mortality temperature) is de temperatuur waarbij er in een bepaalde stad of regio de minste mensen overlijden. Welke temperatuur is de beste voorspeller hiervoor? Bespreek aan de hand van onderstaande spreidingsdiagrammen. MFT staat voor most frequent temperature. (https://www.nature.com/articles/s41467-019-12663-y)




16. Schets een puntenwolk bij onderstaand artikel uit De Morgen (augustus 2023). Benoem de assen en de punten van de puntenwolk

**U bent heel lang, misschien nog altijd, een van de weinige topvrouwen geweest in AI. Hoe komt dat?**

"De grootste instroom naar AI komt nog altijd vanuit de computerwetenschappen, en daar zijn weinig meisjes te vinden. Tenminste, in Europese landen. In Latijns-Amerika ligt die balans anders. Brazilië bijvoorbeeld is traditioneel goed in AI en daar zie je veel meer topvrouwen. Als je in een land woont waar je niet de luxe hebt om iets te gaan studeren wat je graag doet, maar waar je zeker moet zijn van werk, ga je strategischer kiezen voor een opleiding die werkzekerheid biedt. De ingenieursopleidingen trekken daar bijvoorbeeld ook veel vrouwen aan. Blijkbaar is er een negatieve correlatie tussen vrouwen die voor een wiskundige of technische opleiding kiezen en de welvaart van een land."

17. Via https://fbref.com/en/comps/37/2022-2023/2022-2023-Belgian-Pro-League-Stats kan je spreidingsdiagrammen maken van gegevens van de Belgische voetbalcompetitie (bijv goals voor vs puntentotaal, goals tegen vs puntentotaal, doelpuntensaldo vs puntentotaal)